ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК»

На правах рукописи

ШАРАПОВ СОДБО ЖАМБАЛОВИЧ

Полногеномное исследование ассоциаций уровней N-гликозилирования белков плазмы крови человека

1.5.7. - генетика

Диссертация на соискание ученой степени кандидата биологических наук

Научный руководитель:

д.б.н. Юрий Сергеевич Аульченко

Новосибирск 2022

Оглавление

Список сокращений

ПГИА – полногеномное исследование ассоциаций

ОНП – однонуклеотидный полиморфизм

ГДФ - Гуанозиндифосфат

СВЭЖХ (UPHLC) – сверхвысокоэффективная жидкостная хроматография (Ultra performance liquid chromatography)

ВЭЖХ (HPLC) – высокоэффективная жидкостная хроматография (High performance liquid chromatography)

HWE - Hardy Weinberg Equilibrium – равновесие Харди-Вайнберга

ПО - программное обеспечение

П.н. – пар нуклеотидов

м.п.н. – 1,000,000 пар нуклеотидов

т.п.н – 1,000 пар нуклеотидов

QTL – Quantative Trait Locus; локус, ассоциированный с количественным признаком

eQTL – expression Quantative Trait Locus- локус, ассоциированный с уровнем экспрессии определенного гена в определенной ткани или типе клеток

IgG – иммуноглобулин G

ЭР – эндоплазматический ретикулум

АГ – Аппарат Гольджи

Введение

Актуальность темы исследования

Гликозилирование - присоединение углеводного остатка (гликана) - является одной из самых распространенных пост- и ко-трансляционных модификаций белков [1, 2]. Известно, что более 40% (по массе) белков плазмы крови человека гликозилированы [3]. Гликозилирование влияет на биохимические свойства белков (пространственную конфигурацию, фолдинг, растворимость, время полужизни и т.п.) [4–7], и, как следствие, на их биологические функции, включая белок-белковые взаимодействия, взаимодействия белков с рецепторами, клеточные взаимодействия, взаимодействия хозяин-паразит и т.п. [5–9].

В 1980 году было открыто первое врожденное заболевание, вызванное нарушением гликозилирования (англ. CDG - congenital disorder of glycosylation). Данное открытие положило начало исследованию роли гликанов в этиологии заболеваний человека. В результате исследований было показано, что ряд мутаций в генах, кодирующих ферменты биосинтеза гликанов, приводят к серьезным врожденным моногенным заболеваниям человека [10].

Прогресс в области методов анализа гликозилирования белков позволил к началу 2010-х годов проводить исследования популяционной изменчивости N-гликома белков плазмы крови - набора N-гликанов (связанных с белками плазмы крови) и их концентраций - и ее ассоциации с распространенными заболеваниями человека. В качестве исследуемого признака, как правило, выбирался общий N-гликом плазмы крови человека, либо N-гликом иммуноглобулина G (IgG) – наиболее представленного N-гликопротеина в плазме крови человека. Выбор данных признаков обусловлен более простым (по сравнению с N-гликомом других тканей) забором материала и высокой

производительностью существующих методов анализа N-гликома белков плазмы, позволяющих анализировать выборки объемом тысячи и десятки тысяч образцов. В результате масштабных популяционных исследований было показано, что уровни различных N-гликанов белков плазмы крови ассоциированы с риском развития мультифакторных заболеваний человека [11, 12].

Найденные ассоциации уровней N-гликанов белков плазмы крови с риском мультифакторных заболеваний человека стали основанием для рассмотрения N-гликанов в качестве потенциального источника биомаркеров и терапевтических мишеней [8, 13–20]. Однако разработка биомаркеров заболеваний, диагностических и терапевтических мишеней на основе N-гликанов затрудняется неполнотой накопленных знаний о регуляции гликозилирования *in vivo*. Понимание молекулярно-биологических основ регуляции гликозилирования белков позволит не только получить новые знания о регуляции столь важного биологического процесса, но и пролить свет на то, как гликаны вовлечены в контроль мультифакторных заболеваний и признаков человека [21], а также разработать новые биомаркеры, диагностические тесты и лекарственные средства [22, 23].

Гликозилирование белков варьируется в зависимости от ткани и типа клеток [24, 25]. В отличии от транскрипции или трансляции, биосинтез гликанов не является матричным процессом. Биохимическая структура гликанов не закодирована в последовательности ДНК или РНК. Биосинтез гликанов представляет собой разветвленную сеть биохимических реакций [26]. Конечная структура гликана определяется взаимодействием множества молекул и факторов – субстратов и их доступностью, активностью ферментов биосинтеза и деградации гликанов (в первую очередь ферментов из семейств гликозилтрансфераз и гликозидаз[26]), их локализацией и конкуренцией за субстрат, и белками – транспортерами [27–31]. Биохимические пути гликозилирования хорошо изучены [26], однако понимание механизмов

общей и тканеспецифической регуляции этих биохимических реакций *in vivo* ограничено [32]. Это затрудняет как интерпретацию наблюдаемых ассоциаций уровней N-гликанов белков с заболеваниями человека, так и разработку биомаркеров и терапевтических мишеней на их основе.

Генетика и ее методология позволяет изучать сложные биологические системы *in vivo*. В частности, определение генов и регуляторных областей генома, вариация в которых приводит к изменениям в гликоме, может пролить свет на механизмы регуляции гликозилирования *in vivo*. Появление в 2010-х годах крупных выборок людей (более тысячи образцов), для которых было проведено полногеномное генотипирование и определены профили гликозилирования белков плазмы крови, дало возможность проводить исследования регуляции процессов гликозилирования с помощью современных методов генетического анализа. Самым широко используемым методом картирования локусов комплексных признаков и заболеваний человека является полногеномное исследование ассоциаций (ПГИА). В рамках данного метода тестируется ассоциация между исследуемым признаком и большим числом (от сотен тысяч до десятков миллионов) генетических маркеров, равномерно распределенных по геному. При этом, как правило, анализируются выборки размером от нескольких тысяч до миллионов особей. Результатом ПГИА является набор геномных локусов, замены в которых приводят к изменению исследуемого признака.

Знание локуса является отправной точкой в исследовании его роли в формировании исследуемого признака. Поскольку каждый локус может содержать множество полиморфных участков и кодирующие участки большого числа генов, спектр возможных гипотез о функциональных вариантах и кандидатных генах может быть огромным. Проверка каждой из гипотез в *in vitro*, и тем более, *in vivo* экспериментах превращается в непосильную задачу. Поэтому после проведения ПГИА принято проводить более быстрое функциональное исследование *in silico* для приоритизации

гипотез о возможных биологических механизмах, лежащих в основе найденных ассоциаций и, как следствие, для проведения более оптимальных исследований *in vitro* и *in vivo*.

К настоящему времени было проведено два ПГИА уровней N-гликанов белков плазмы крови человека [33, 34], в которых измерение уровней N-гликанов белков плазмы крови производилось с помощью технологии высокоэффективной жидкостной хроматографии (ВЭЖХ). В результате была найдена ассоциация шести локусов. Один из них расположен на 12-ой хромосоме и содержит гены *SPPL3*, *C12orf43*, *OASL* и *HNF1A−AS1/HNF1A*. Роль этого локуса в контроле N-гликозилирования белков ранее не была известна. Данный локус показал ассоциацию с уровнем фукозилирования белков плазмы крови. Ген *HNF1A* кодирует фактор транскрипции гепатоцитов и экспрессируется в органах эндодермального происхождения – печени, почках, поджелудочной железе и т.д. Принимая во внимание, что гепатоциты – клетки печени – являются одним из главных источников гликопротеинов в плазме крови, исследователи сформулировали гипотезу о возможном влиянии *HNF1A* на экспрессию фукозилтрансфераз. Данная гипотеза была подтверждена в последующем функциональном исследовании [33], которое показало, что продукт гена *HNF1A* регулирует экспрессию генов фукозилтрансфераз (*FUT3, FUT5, FUT6, FUT8, FUT10, FUT11*) в клетках линии HepG2, полученной из гепатоцитов. Более того, было показано, что *HNF1A* регулирует экспрессию ферментов, необходимых для синтеза ГДФ-фукозы – субстрата для фукозилтрансфераз. В результате была показана важная роль гена *HNF1A* в контроле фукозилирования белков плазмы крови. Таким образом, исследование генетического контроля уровней N-гликанов плазмы крови человека методом ПГИА позволяет формировать ранее неизвестные гипотезы о регуляции данного процесса. Стоит отметить, что еще в 1996 году было обнаружено, что мутации в гене *HNF1A* вызывают сахарный диабет взрослого типа у молодых (Maturity Onset Diabetes of the Young 3, MODY-3) [35]. Основываясь на результатах ПГИА гликома плазмы [33, 34],

показавших роль гена *HNF1A* в контроле гликома плазмы крови, были обнаружены потенциальные гликомные биомаркеры заболевания MODY-3 [20], показавших высокую диагностическую точность. Это подтверждает актуальность исследования генетического контроля гликозилирования для разработок методов прогнозирования, диагностирования, профилактики и лечения заболеваний человека.

За время, прошедшее после публикации последнего ПГИА N-гликома белков плазмы в 2011 году, появились новые технологии измерения уровней гликанов [36–39], а также новые референтные выборки для импутирования геномных данных (такие как 1000 геномов [40], HRC [41], и TOPMed [42]), разрешение которых на порядок больше таковых, доступных в 2011 году (например HapMap [43, 44]). Среди современных технологий анализа уровней гликозилирования белков плазмы крови человека, наибольшее распространение получил метод сверхвысокоэффективной жидкостной хроматографии (СВЭЖХ) [39]. Разнообразие гликанов, уровни которых измеряются методом СВЭЖХ, как и точность измерения, выше, чем у метода ВЭЖХ. Генетический анализ уровней расширенного набора гликанов, измеренных с большей точностью, позволил увеличить мощность анализа генетических ассоциации, и, как следствие, более точно охарактеризовать регуляцию процессов гликозилирования найденными локусами.

Однако проведение ПГИА N-гликома плазмы крови, измеренного технологией СВЭЖХ было невозможным из-за того, что в зависимости от исследования, число признаков варьирует от 36 до 42 [39, 45–47]. Данные различия обусловлены изменениями в протоколах анализа СВЭЖХ, как на этапах проведения хроматографического анализа (ряд пиков могут оказаться недостаточно разделенными), так и на этапе определения границ пиков – интеграции. Проведение полногеномного исследования ассоциаций на материале нескольких выборок возможно только при условии того, что в каждой из выборок анализ ассоциаций проводится для единого

(гармонизированного) набора признаков. Таким образом, для проведения ПГИА гликома плазмы крови человека с последующим подтверждением результатов на независимых выборках, требуется разработка и применение метода гармонизации гликомных профилей СВЭЖХ, полученных в анализируемых выборках.

Принимая во внимание ограниченное число работ, посвященных исследованию генетического контроля гликозилирования, в том числе из-за недостаточного методологического обеспечения, исследование генетического контроля гликозилирования белков плазмы крови человека с использованием современных методов измерения гликома и генетических данных высокого разрешения является актуальной проблемой современной генетики человека.

Степень разработанности темы исследования

На данный момент опубликовано два ПГИА уровней N-гликанов белков плазмы крови человека [33, 34]. В данных работах использовались устаревшая на данный момент технологии профилирования N-гликома – ВЭЖХ – и генетические данные, импутированные с использованием гаплотипов выборки HapMap2 [44]. С момента публикации последнего ПГИА N-гликома плазмы в 2011 году, появились новые технологии измерения уровней гликанов [36–39], а также новые референтные выборки для импутирования генетических данных (такие как 1000 геномов [40], HRC [41], и TOPMed [42]), разрешение которых на порядок больше таковых, доступных в 2011 году (например HapMap [43, 44]). Среди современных технологий анализа уровней гликозилирования белков плазмы крови человека, наибольшее распространение получил метод сверхвысокоэффективной жидкостной хроматографии (СВЭЖХ) [39]. Число гликанов, измеряемых методом СВЭЖХ, как и точность их измерения, выше, чем у метода ВЭЖХ. Генетический анализ уровней расширенного набора гликанов, измеренных с большей точностью, позволил увеличить мощность анализа генетических ассоциаций, и, как следствие, более точно

охарактеризовать регуляцию процессов гликозилирования найденными локусами.

Цели и задачи

Целью данной работы является поиск генов, участвующих в контроле N-гликозилирования белков плазмы крови человека. Для достижения этой цели были поставлены следующие задачи:

1. Провести полногеномное исследование ассоциаций генетических маркеров с уровнями N-гликанов белков плазмы крови человека на материале выборки TwinsUK.
2. Разработать и валидировать метод гармонизации данных об уровнях N-гликанов белков плазмы крови человека, необходимый для сопоставления результатов, полученных в различных выборках, и применить его для гармонизации данных независимых выборок.
3. Подтвердить результаты анализа ассоциаций на материале независимых выборок.
4. Приоритизировать в найденных локусах гены-кандидаты, наиболее вероятно влияющие на уровни N-гликанов белков плазмы крови человека.

Научная новизна

В данной работе впервые проведено полногеномное исследование ассоциации уровней N-гликанов белков плазмы крови человека, измеренных технологией сверхвысокоэффективной жидкостной хроматографии (СВЭЖХ). Для проведения исследования на материале нескольких выборок впервые был разработан метод гармонизации гликомных профилей, измеренных технологией СВЭЖХ.

Впервые были обнаружены ассоциации 10 локусов с уровнями гликозилирования белков плазмы крови, для 9 из которых ассоциация была

подтверждена на независимых выборках. Также была подтверждена ассоциация 6 локусов, найденных в предыдущих исследованиях. В общей сложности, была подтверждена ассоциация 15 локусов.

На основе результатов функционального исследования *in silico*, для 15 локусов были предложены гены – кандидаты, наиболее вероятно участвующие в регуляции процессов гликозилирования. Впервые показана возможная роль генов *RUNX3, IKZF1, SMARCB1, DERL3, CHCHD10, IGH, TMEM121* и *HLA* в генетическом контроле уровней N-гликозилирования белков плазмы крови человека.

Теоретическая и научно-практическая ценность

Результаты данной работы расширяют представление о генетической регуляции N-гликозилирования белков плазмы крови человека. Полногеномное исследование ассоциации уровней гликозилирования белков плазмы крови человека, измеренных технологией СВЭЖХ, позволило установить роль пятнадцати локусов в регуляции процессов гликозилирования. Биоинформатический анализ найденных локусов расширил наше представление о возможных генах и процессах, участвующих в регуляции N-гликозилирования белков. Роль одного из предложенных генов-кандидатов – транскрипционного фактора *IKZF1* в регуляции фукозилирования белков – была доказана в результате *in vitro* эксперимента, что подкрепляет научную состоятельность сформулированных гипотез о генах-кандидатах.

Получение и публикация в открытом доступе суммарных статистик генетических ассоциаций для 113 N-гликомных признаков позволит всем заинтересованным исследователям использовать эти данные для совместного анализа с интересующими их данными функциональной геномики и количественной генетики. В частности, полученные результаты ПГИА гликома плазмы крови будут востребованы в исследованиях по поиску

биомаркеров и терапевтических мишеней гликом-ассоциированных заболеваний.

Разработанный и успешно примененный протокол гармонизации гликомных профилей, измеренных технологией СВЭЖХ, будет востребован как в исследованиях генетического контроля гликозилирования, так и в эпидемиологических исследованиях связи гликозилирования с риском развития заболеваний человека. При этом данный метод может применяться не только в исследованиях общего гликозилирования плазмы крови, но и в исследованиях гликозилирования конкретных белков – иммуноглобулинов, трансферринов и т.п.

Методология и методы диссертационного исследования

В данной работе использовались первичные результаты измерения уровней гликозилирования белков плазмы крови человека и генотипы испытуемых. Первичная обработка данных СВЭЖХ и их контроль качества для образцов исследуемых выборок проводились согласно принятым в данной области стандартам и включали в себя гармонизацию, логарифмирование, нормализацию, коррекцию сдвига систематической ошибки измерения в разных сериях, определение образцов – статистических выбросов и расчет производных признаков. Контроль качества генотипов испытуемых проводился согласно принятым и опубликованным стандартам [48].

Картирование геномных локусов, вариация в которых ассоциирована с изменением уровней N-гликанов плазмы крови человека, проводилось с помощью метода полногеномного анализа ассоциации. Контроль качества суммарных статистик генетических ассоциаций и их мета-анализ проводился согласно принятым и опубликованным стандартам [49].

Биоинформатический анализ полученных результатов проводился с использованием данных функциональной геномики – GTEx [50], CEDAR [51], и с использованием методов приоритизации функциональных вариантов –

VEP [52], FATHMM-XF [53], FATHMM-InDel [54], и генов - SMR/HEIDI [55] и DEPICT [56].

Степень достоверности результатов

Степень достоверности результатов подтверждается согласованностью результатов полногеномного анализа ассоциации, выполненного в данной работе, с опубликованными ранее результатами. Полученные результаты также были подтверждены на материале нескольких независимых выборок, набранных в различных популяциях людей, что говорит о высокой степени достоверности полученных результатов и их обобщающей способности (генерализуемости).

Положения, выносимые на защиту

1. Популяционная изменчивость уровней N-гликанов, связанных с белками плазмы крови человека, контролируется как минимум 15 локусами генома, 9 из которых определены впервые.
2. Генами-кандидатами, вовлеченными в процесс N-гликозилирования, являются гены регуляторов транскрипции (*IKZF1*, *SMARCB1* и *RUNX3*), деградации гликопротеинов (*DERL3*), тяжелой цепи иммуноглобулинов (*IGH*) и гены с неизвестной функцией (*TMEM121* и *CHCHD10*).

Структура и объём работы

Работа состоит из введения, обзора литературы, материалов и методов, результатов, обсуждения, заключения, выводов, списка литературы (167 источников) и приложений. Объем работы составляет 153 страницы. Работа включает 10 таблиц, 17 рисунков и 4 таблицы в приложении.

Личный вклад автора

Цели и задачи были сформулированы автором совместно с научным руководителем. Основные результаты, изложенные в диссертации, получены

и проанализированы автором лично. Материалы для исследования - первичные данные измерения уровней гликозилирования белков плазмы крови человека и генотипы испытуемых - были любезно предоставлены хорватскими (Genos Ltd., PainOmics-St. Catherine), английскими (TwinsUK, SABRE, PainOmics-UK), немецкими (EPIC-Potsdam), итальянскими (PainOmics-UNIPR), бельгийскими (PainOmics-ZOL) и шотландскими (SOCCS) коллегами в рамках научной коллаборации.

Апробация результатов

Материалы настоящей работы вошли в отчеты по гранту Российского Научного Фонда № 19-15-00115 «Гены - регуляторы гликозилирования белков человека». Результаты работы были представлены лично автором на 9 международных научных конференциях в виде 7 устных и 2 стендовых докладов:

1. Sodbo Sharapov, Yakov Tsepilov, Elizaveta Elgaeva, Evgeny Tiys, Arina Nostaeva, Frano Vuckovic, Irena Trbojević-Akmačić, Michel Georges, Karsten Suhre, Nishi Chaturvedi, Harry Campbell, Malcolm Dunlop, Frances Williams, Matthias B. Schulze, Tim Spector, Gordan Lauc, Yurii S. Aulchenko. Meta-analysis of genome-wide association studies for N-glycosylation in 10,000 individuals. European Human Genetics Conference 2021 (ESHG 2021), Vienna, Austria, 2021;
2. Sharapov S., Tsepilov Y., Elgaeva E., Tiys E., Nostaeva A., Vuckovic F., Trbojević-Akmačič I., Georges M., Suhre K., Chaturvedi N., Campbell H., Dunlop M., Williams F., Schulze M., Spector T., Lauc G., Aulchenko Y. Mapping genes involved in control of N-glycosylation of blood glycoproteins through a large genome-wide association study. MCCMB'21. 10th Moscow Conference on Computational Molecular Biology, Moscow, Russia, 2021; 20;
3. Sharapov S. “Genome-wide association study identifies tissue-specific regulation of human protein N-glycosylation”. International Conference

on Quantitative Genetics (ICQG6), Brisbane, Australia, November 2020; 243;

4. Sharapov S., Feoktistova S., Klaric L., Campbell H., Schulze M., Aulchenko Yu., Tsepilov Ya. A., Tiys E., Suhre K., Dunlop M., Spector T., Elgaeva E.E., Vuckovic F., Williams F., Lauc G., Chaturvedi N. “Results of genome-wide association study of plasma proteome N-glycosylation in 10,000 samples”. BGRS/SB-2020: 12th International Multiconference “Bioinformatics of Genome Regulation and Structure/Systems Biology”. Novosibirsk, Russia, 2020; 103-104;
5. Шарапов С.Ж., Цепилов Я.А., Лауц Г., Аульченко Ю.С. Генетическая регуляция N-гликозилирования белков плазмы крови человека. «Съезд Биохимиков России", Сочи, Россия, 2019; 143;
6. Sodbo Zh. Sharapov, Yakov A. Tsepilov, Lucija Klaric, Massimo Mangino, Gaurav Thareja, Alexandra S. Shadrina, Mirna Simurina, Concetta Dagostino, Julia Dmitrieva, Marija Vilaj, Frano Vuckovic, Tamara Pavic, Jerko Stambuk, Irena Trbojevic-Akmacic, Jasminka Kristic, Jelena Simunovic, Ana Momcilovic, Harry Campbell, Margaret Doherty, Malcolm G Dunlop, Susan M Farrington, Maja Pucic-Bakovic, Christian Gieger, Massimo Allegri, Edouard Louis, Michel Georges, Karsten Suhre, Tim Spector, Frances MK Williams, Gordan Lauc, Yurii Aulchenko. Genome-wide association study finds new loci affecting N-glycosylation of human blood plasma proteins. MCCMB'19. 9th Moscow Conference on Computational Molecular Biology, Moscow, Russia, 2019; 27;
7. Sharapov S., Tsepilov Y., Klaric L., Mangino M., Thareja G., Shadrina A., Simurina M., Dagostino C., Dmitrieva J., Vilaj M., Vuckovic F., Pavic T., Stambuk E., Trbojevic-Akmacic I., Kristic J., Simunovic J., Momcilovic A., Campbell H., Doherty M., Dunlop M., Farrington S., Pucic-Bakovic M., Gieger C., Allegri M., Louis E., Georges M., Suhre K., Spector T., Williams F., Lauc G., Aulchenko Y. Genetic loci, associated with

glycosylation of human total plasma proteins: a 2019 update. ISABS Conference on Forensic and Anthropologic Genetics, Split, Croatia, 2019;

8. Sharapov S., Tsepilov Y., Klaric L., Mangino M., Vuckovic F., Campbell H., Dunlop M., Farrington S., Suhre K., Spector T., Lauc G., Aulchenko Y. Defining the genetic control of human blood plasma glycome using genome-wide association study. BGRS\SB-2018. Bioinformatics of Genome Regulation and Structure\Systems Biology, Novosibirsk, Russia, 2018; 130;
9. Sharapov S., Tsepilov Y., Klaric L., Mangino M., Vuckovic F., Campbell H., Dunlop M., Farrington S., Gaurav T., Suhre K., Spector T., Lauc G., Aulchenko Y. Understanding the genetic control of human blood plasma glycome using genome-wide association study. 2nd GlycoCom, Dubrovnik, Croatia, 2018; 81-82.

Публикации по теме работы

По материалам диссертации опубликовано 4 научные работы в изданиях, индексируемых в базах данных «Скопус» (Scopus) и «Сеть науки» (Web of Science):

1. **Sodbo Zh Sharapov**, Alexandra S Shadrina, Yakov A Tsepilov, Elizaveta E Elgaeva, Evgeny S Tiys, Sofya G Feoktistova, Olga O Zaytseva, Frano Vuckovic, Rafael Cuadrat, Susanne Jäger, Clemens Wittenbecher, Lennart C Karssen, Maria Timofeeva, Therese Tillin, Irena Trbojević-Akmačić, Tamara Štambuk, Najda Rudman, Jasminka Krištić, Jelena Šimunović, Ana Momčilović, Marija Vilaj, Julija Jurić, Anita Slana, Ivan Gudelj, Thomas Klarić, Livia Puljak, Andrea Skelin, Antonia Jeličić Kadić, Jan Van Zundert, Nishi Chaturvedi, Harry Campbell, Malcolm Dunlop, Susan M Farrington, Margaret Doherty, Concetta Dagostino, Christian Gieger, Massimo Allegri, Frances Williams, Matthias B Schulze, Gordan Lauc, Yurii S Aulchenko (2021). Replication of 15 loci involved in human

plasma protein N-glycosylation in 4802 samples from four cohorts. Glycobiology, 31(2), 82–88, https://doi.org/10.1093/glycob/cwaa053;

2. Lucija Klarić, Yakov A Tsepilov, Chloe M Stanton, Massimo Mangino, Timo Tõnis Sikka, Tõnu Esko, Eugene Pakhomov, Perttu Salo, Joris Deelen, Stuart J McGurnaghan, Toma Keser, Frano Vučković, Ivo Ugrina, Jasminka Krištić, Ivan Gudelj, Jerko Štambuk, Rosina Plomp, Maja Pučić-Baković, Tamara Pavić, Marija Vilaj, Irena Trbojević-Akmačić, Camilla Drake, Paula Dobrinić, Jelena Mlinarec, Barbara Jelušić, Anne Richmond, Maria Timofeeva, Alexander K Grishchenko, Julia Dmitrieva, Mairead L Bermingham, **Sodbo Zh Sharapov**, Susan M Farrington, Evropi Theodoratou, Hae-Won Uh, Marian Beekman, Eline P Slagboom, Edouard Louis, Michel Georges, Manfred Wuhrer, Helen M Colhoun, Malcolm G Dunlop, Markus Perola, Krista Fischer, Ozren Polasek, Harry Campbell, Igor Rudan, James F Wilson, Vlatka Zoldoš, Veronique Vitart, Tim Spector, Yurii S Aulchenko, Gordan Lauc, Caroline Hayward. (2020). Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. Science Advances, 6(8), eaax0301. https://doi.org/10.1126/sciadv.aax0301;
3. **Sodbo Zh Sharapov**, Yakov A Tsepilov, Lucija Klaric, Massimo Mangino, Gaurav Thareja, Alexandra S Shadrina, Mirna Simurina, Concetta Dagostino, Julia Dmitrieva, Marija Vilaj, Frano Vuckovic, Tamara Pavic, Jerko Stambuk, Irena Trbojevic-Akmacic, Jasminka Kristic, Jelena Simunovic, Ana Momcilovic, Harry Campbell, Margaret Doherty, Malcolm G Dunlop, Susan M Farrington, Maja Pucic-Bakovic, Christian Gieger, Massimo Allegri, Edouard Louis, Michel Georges, Karsten Suhre, Tim Spector, Frances MK Williams, Gordan Lauc, Yurii S Aulchenko (2019). Defining the genetic control of human blood plasma N-glycome using genome-wide association study. Human Molecular Genetics, 28(12), 2062–2077. https://doi.org/10.1093/hmg/ddz054;

4. Xia Shen, Lucija Klarić, **Sodbo Zh Sharapov**, Massimo Mangino, Zheng Ning, Di Wu, Irena Trbojević-Akmačić, Maja Pučić-Baković, Igor Rudan, Ozren Polašek, Caroline Hayward, Timothy D Spector, James F Wilson, Gordan Lauc, Yurii S Aulchenko (2017). Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation. Nature Communications, 8(1), 447. https://doi.org/10.1038/s41467-017-00453-3.

Глава 1. Обзор литературы

1.1. Гликомика – раздел гликобиологии

Гликобиология – наука, изучающая структуру, биосинтез, биологическую роль и эволюцию углеводов - гликанов [57]. Клетки всех без исключения организмов и огромное число макромолекул имеют в своем составе ковалентно присоединенные углеводные остатки [57]. Гликаны могут ковалентно присоединяться к белкам и липидам путем образования гликозидной связи, образуя соответственно гликопротеины и гликолипиды. Гликозилирование является одной из самых распространенных [2] и важных пост- и ко-трансляционных модификаций белков. Известно, что в природе 20% всех видов белков гликозилированы [1]. При этом, более 40% (по массе) всех белков плазмы крови человека N-гликозилированы [3]. Гликозилирование влияет как на физико-химические свойства белков - растворимость, пространственную конфигурацию, фолдинг и т.п. [4–6], так и на их биологическую функцию. Гликопротеины и гликолипиды, расположенные на поверхности клеточных мембран, участвуют в процессах клеточных взаимодействий – между клетками, клетками и внеклеточным матриксом и между клетками и макромолекулами, а также взаимодействий между организмами (хозяин-паразит, симбионт-симбионт и т.п.) [5, 6, 8, 9], тем самым играя огромную роль в развитии и функционировании многоклеточных организмов [58].

Множество исследований химической структуры гликанов и их метаболизма были проведены в первой половине XX века. В то время гликаны рассматривались в основном как структурные элементы и источники энергии в живых системам [57]. Бурное развитие химических, физических и молекулярно-биологических методов исследования гликанов привело к появлению нового раздела молекулярной биологии – гликобиологии. В настоящее время гликобиология является быстро развивающейся областью

научных исследований, имеющей большое значение для многих областей фундаментальных наук, включая биомедицину и биотехнологии [57, 59]. Данная область включает в себя исследования химических и физических свойств гликанов, энзимологии процессов синтеза и деградации гликанов, механизмов распознавания гликанов белками, роли гликанов в функционировании биологических систем, их роли в развитии заболеваний и признаков человека и разработке методов лечения, профилактики, диагностики и прогнозирования заболеваний.

По аналогии с геномикой, транскриптомикой, протеомикой и метаболомикой, гликомика представляет собой систематическое изучение гликома – совокупности всех гликанов и их концентраций в конкретном образце – клеточной культуре, ткани, органе или целом организме. Мономеры, входящие в структуру гликанов, могут образовывать различные гликозидные связи между собой, за счет чего достигается большое разнообразие гликанов, образующих гликом. Более того, гликаны могут присоединяться к различным сайтам связывания на макромолекулах (белкам, липидам), образуя гликоконъюгаты – углеводсодержащие биополимеры. Помимо структурного разнообразия, гликом может изменяться в ходе развития, дифференциации клеток [60], старения [61], воспаления [62], заболеваний [14, 21, 63–66] и т.д. Далее по тексту под N-гликомом будет пониматься набор N-гликанов, связанных с белками, и их относительных концентраций.

Технологический прогресс в области методов анализа N-гликома позволил к началу 2010-ых годов проводить когортные исследования ассоциаций гликома с заболеваниями и признаками человека. На данный момент показана ассоциация N-гликанов с риском развития целого рядя мультифакторных заболеваний человека [12, 67, 68], например диабетом первого и второго типа [20, 69, 70], ревматоидным артритом [21], болезнью Паркинсона [65], воспалительным заболеванием кишечника [71, 72], дорсалгией [63], гипертензией [73], сердечно-сосудистыми [14] и

онкологическими [13, 45, 74, 75] заболеваниями. Список заболеваний, ассоциированных с N-гликомом, продолжает расширяться. Стоит отметить, что изменения в N-гликозилировании белков играют большую роль в этиологии аутоиммунных заболеваний [18]. Не менее интересной является связь N-гликанов с биологическим возрастом человека. В работах [46, 61, 76] показана зависимость между N-гликанами белков плазмы крови и хронологическим возрастом человека.

Результаты обсервационных исследований ассоциации гликозилирования белков с риском развития заболеваний не дают ответа на вопрос о причинно-следственных связях между ними и, тем более, о молекулярно-биологических механизмах, лежащих в основе данных связей. Начиная с 1983 года проводятся исследования функциональных последствий изменений в гликозилировании белков [62, 77]. Иммуноглобулины класса G являются наиболее изученными с этой точки зрения ввиду важности данного гликопротеина в адаптивном иммунитете. Молекула IgG имеет консервативный сайт гликозилирования Asn297, находящийся в консервативном домене CH2 тяжелой цепи. Данный домен играет важную роль в связывании с Fcγ-рецепторами, что в свою очередь влияет на эффекторную функцию IgG. Было показано, что снижение уровня фукозилирования IgG усиливает антителозависимую клеточную цитотоксичность (antibody-dependent cellular cytotoxicity, ADCC) [78]. Дальнейшие исследования с использованием методов кристаллографии показали, что отсутствие фукозилирования IgG ведет к увеличению аффинности между Fc доменом IgG и рецептором FcγRIIIA, поскольку наличие фукозы в составе гликана IgG ведет к появлению стерических затруднений при взаимодействии [79].

На основание всего вышесказанного можно сделать вывод о фундаментальной важности изучения гликома для решения задач диагностики, предсказания профилактики и лечения заболеваний человека. В

2012 году Национальная Академия Наук США представила доклад о необходимости полномасштабного изучения гликома, так как «…гликаны напрямую вовлечены в патогенез практически всех известных заболеваний…» (https://www.nap.edu/catalog/13446/transforming-glycoscience-a-roadmap-for-the-future).

1.2. Строение и разнообразие гликанов

Углеводы – одна из основных групп макромолекул, выделяемых в биологии наряду с белками, липидами и нуклеиновыми кислотами. Благодаря способности к полимеризации и наличию высокого числа хиральных атомов, моносахариды могут образовывать огромное количество различных стерео- и региоизомеров. По степени полимеризации различают четыре основных группы углеводов: моносахариды (глюкоза, фруктоза, галактоза т.д.), дисахариды (молекулы, состоящие из двух моносахаридов, связанных гликозидной, напр. сахароза, лактоза, мальтоза), полисахариды с повторяющимися звеньями, образующие как линейные, так и разветвленные соединения (О-антигены бактерий, амилоза, целлюлоза, хитин) и гликаны - сложные олигосахариды с неповторяющимися звеньями, свободные или связанные с белками или липидами (гликопротеины, протеогликаны и гликолипиды).

В свою очередь, гликаны, связанные с белками, подразделяются на N-гликаны, O-гликаны и C-гликаны. N-гликаны образуют гликозидную связь с атомом азота (N), находящимся в составе аминокислоты аспарагина (см. Рис. 1). O-гликаны образуют гликозидную связь с гидроксилом аминокислот серина или треонина. C-гликаны образуют гликозидную связь с атомом углерода, входящим в состав аминокислоты триптофана. C-гликозилирование встречается редко по сравнению с N- и O-гликозилированием [80].

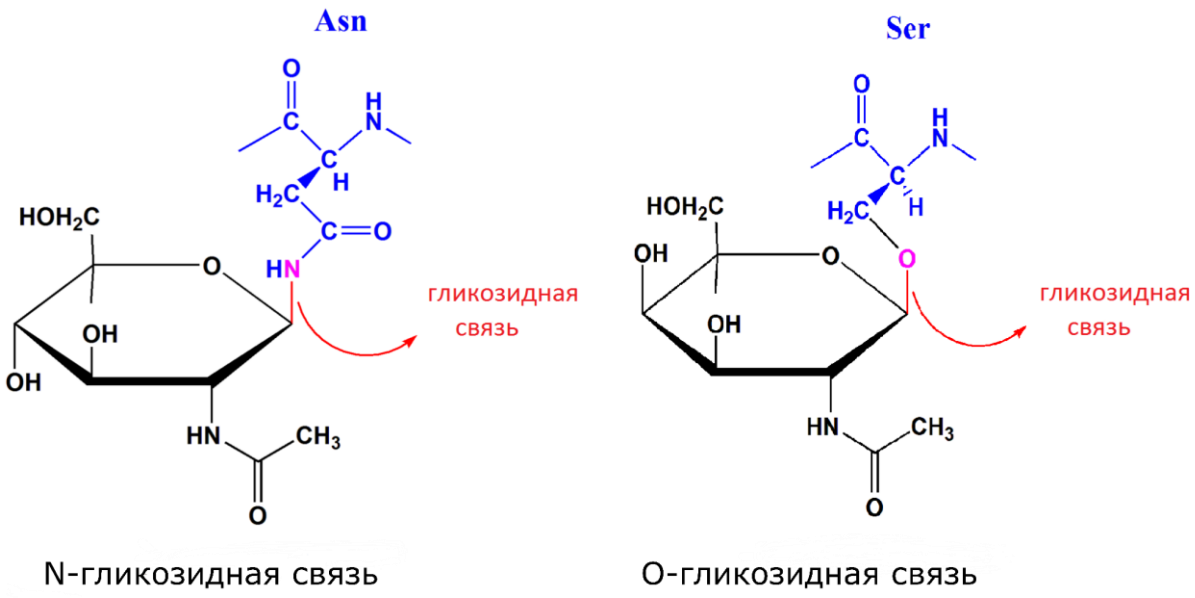


Рис. 1. Схема образования гликозидной связи углевода и аминокислоты. Рисунок слева – образование N-гликозидной связи, рисунок справа – образование O-гликозидной связи. Синим цветом показаны структуры аминокислот, входящих в состав полипептида. Черным цветом показана структура моносахарида, входящего в состав гликана. (Адаптировано из https://en.wikibooks.org/wiki/Structural_Biochemistry/Carbohydrates с изменениями).

Важным отличием N-гликозилирования от O-гликозилирования является тот факт, что N-гликозидная связь образуется только с аминокислотой аспарагином, входящей в состав мотива Asn-X-Ser/Thr, где в качества «X» может выступать любая аминокислота за исключением пролина. Благодаря этому отличию, обработка образца, содержащего N-гликозилированный белок, ферментом PNGase F позволяет специфично расщепить N-гликозидную связь, и тем самым, высвободить N-гликаны в раствор для дальнейшего анализа [81]. В отличии от N-гликозилирования, сайт O-гликозилирования не имеет консенсусной последовательности и существующие методы выделения O-гликанов (например бета-элиминирование) обладают меньшей специфичностью по сравнению с таковым для N-гликанов [82]. Это является одной из причин того, что на данный момент наиболее хорошо разработаны технологии и протоколы определения структуры и анализа уровней N-гликанов.

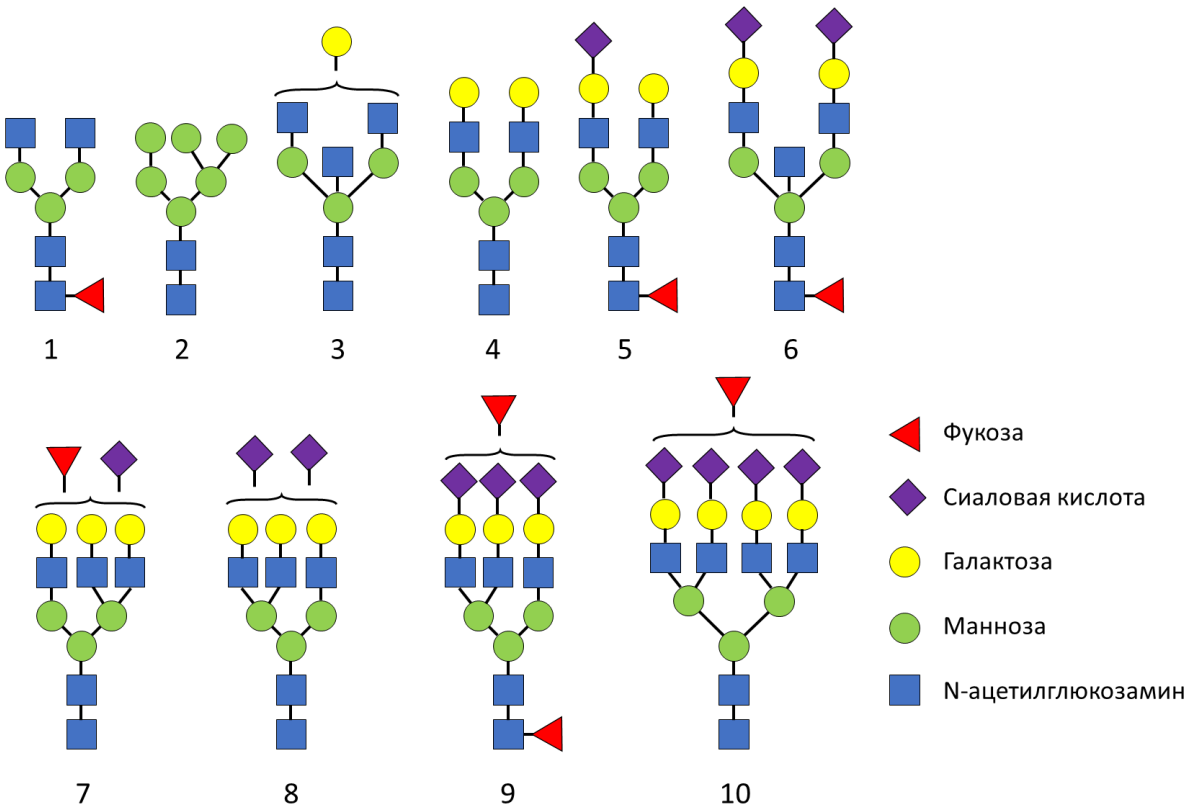


Рис. 2. Примеры структуры N-гликанов. Структура 2 – гликан с большим числом остатков маннозы, остальные гликаны имеют комплексную структуру. Структуры 1,3, 4-6: биантеннарные гликаны, 7-9 – триантеннарные гликаны, структура 10 – тетра-антеннарный гликан; у структур 1, 5, 6 и 9 присутствует фукозилирование остова N-гликана; структуры 7, 9 и 10 антеннарно фукозилированы (фукоза присоединена к антенне). В структурах 3 и 6 присутствует бисектный (bisecting) остаток N-ацетилглюкозамина (GlcNAc). В структурах 3-10 и 5-10 присутствуют остатки галактозы и сиаловой кислоты соответственно.

Наиболее распространенными мономерами в составе N-гликанов являются моносахариды манноза, фукоза, галактоза, N-ацетилглюкозамин (GlcNAc) и сиаловая кислота (см. Рис. 2). В структуре N-гликанов всегда присутствует остов (Manα1-3(Manα1-6)Manβ1-4GlcNAcβ1–4GlcNAcβ1–Asn-X-Ser/Thr), к которому присоединяются остальные мономеры путем образования гликозидной связи. Стрsуктура N-гликанов может быть разветвленной и включать в себя одну (редко) или несколько (как правило) ветвей – антенн. К каждой из антенн могут присоединяться мономеры – галактоза, сиаловая кислота или фукоза. Фукоза также может присоединяться непосредственно к N-ацетилглюкозамину в составе остова N-гликана. Отрицательно заряженная сиаловая кислота – единственный мономер в составе N-гликанов, который несет заряд. Гликаны, не несущие в своем составе сиаловой кислоты, являются нейтральными.

Таким образом, согласно биохимической структуре, N-гликаны можно классифицировать следующим образом:

1. Наличие или отсутствие рассечения остова (bisecting);
2. Ветвление сахарной цепи - ди/три/тетра антеннарные гликаны;
3. Наличие или отсутствие фукозилирования GlcNAc в составе остова N-гликана (core fucosylation);
4. Наличие или отсутствие фукозилирования антеннарных цепей;
5. Галактозилирование антеннарных цепей: отсутствие галактозилирования /моно /ди /три /тетра галактозилирование;
6. Сиалирование антеннарных цепей - нейтральные гликаны /моно /ди /три /тетра сиалирование.

На основе данной классификации гликанов можно описать биохимическую структуру гликанов в упрощенном текстовом виде. Одна из наиболее популярных в настоящее время так называемая «Оксофордская номенклатура» гликанов [83] следует следующим правилам:

1. Символ “F” в самом начале обозначения говорит о наличии фукозы, присоединенной к остову;
2. Далее следует последовательность “AN”, где “N” – число антенн (ветвей) в структуре гликана;
3. Далее, в случае наличия рассечения остова, добавляется символ “B” – bisecting;
4. Далее, в случае фукозилирования антеннарных ветвей добавляется символ “F”;
5. Если в структуре гликана присутствует галактоза, присоединенная к одной или нескольким антеннам, то далее следует последовательность “G[n1,n2,...]N” где “N” – число галактоз в составе гликана, а “n1” – обозначает атом углерода в составе галактозы, с которым образована гликозидная связь;

6. Если в структуре гликана присутствует сиаловая кислота, присоединенная к одной или нескольким антеннам, то далее следует последовательность “S[n1,n2,...]N” где “N” – число остатков сиаловой кислоты в составе гликана, а “n1” – обозначает атом углерода в составе сиаловой кислоты, с которым образована гликозидная связь.

Например, обозначение “FA2” говорит о том, что в структуре гликана присутствует остаток фукозы, присоединенный к остову, а также две антенны. Обозначение “A3BG3S1” говорит о том, что в структуре гликана присутствуют три антенны, рассечение остова, галактозилирование трех антенн и сиалирование одной антенны.

1.3. Биосинтез N-гликанов

В отличие от транскриптов и белков, закодированных в последовательности геномной ДНК и синтезирующихся в результате матричных процессов, структура гликанов не закодирована в геноме. Биосинтез гликанов представляет собой разветвленную сеть биохимических реакций [26]. Конечная структура гликана определяется взаимодействием множества молекул и факторов – субстратов и их доступностью, активностью ферментов биосинтеза и деградации гликанов, их локализацией и конкуренцией за субстрат, и белками – транспортерами [27–30]. Было также показано, что структура и разнообразие N-гликанов, присутствующих в определенных клетках или тканях, отчасти регулируется на уровне транскрипции генов, кодирующих белки, которые участвуют в синтезе и деградации гликанов [29, 30, 84].

Биосинтез гликанов происходит в эндоплазматическом ретикулуме (ЭР) и аппарате Гольджи (АГ) и регулируется доступностью субстратов, активностью ферментов, уровнем транскрипции генов, кодирующих ферменты и транспортеры, а также локализацией ферментов внутри органелл

[12]. На данный момент в базе данных KEGG содержится информация о более чем 300 ферментах, участвующих в процессах синтеза и деградации гликанов [85]. Одними из основных ферментов, непосредственно участвующих в биосинтезе N-гликанов, являются гликозилтрансферазы - ферменты, которые переносят активированные моносахариды к растущему гликану. В специализированной базе данных CAZy [26] содержится аннотация и классификация более 200 гликозилтрансфераз, по крайней мере 40 из которых относятся к пути N-гликозилирования белков.

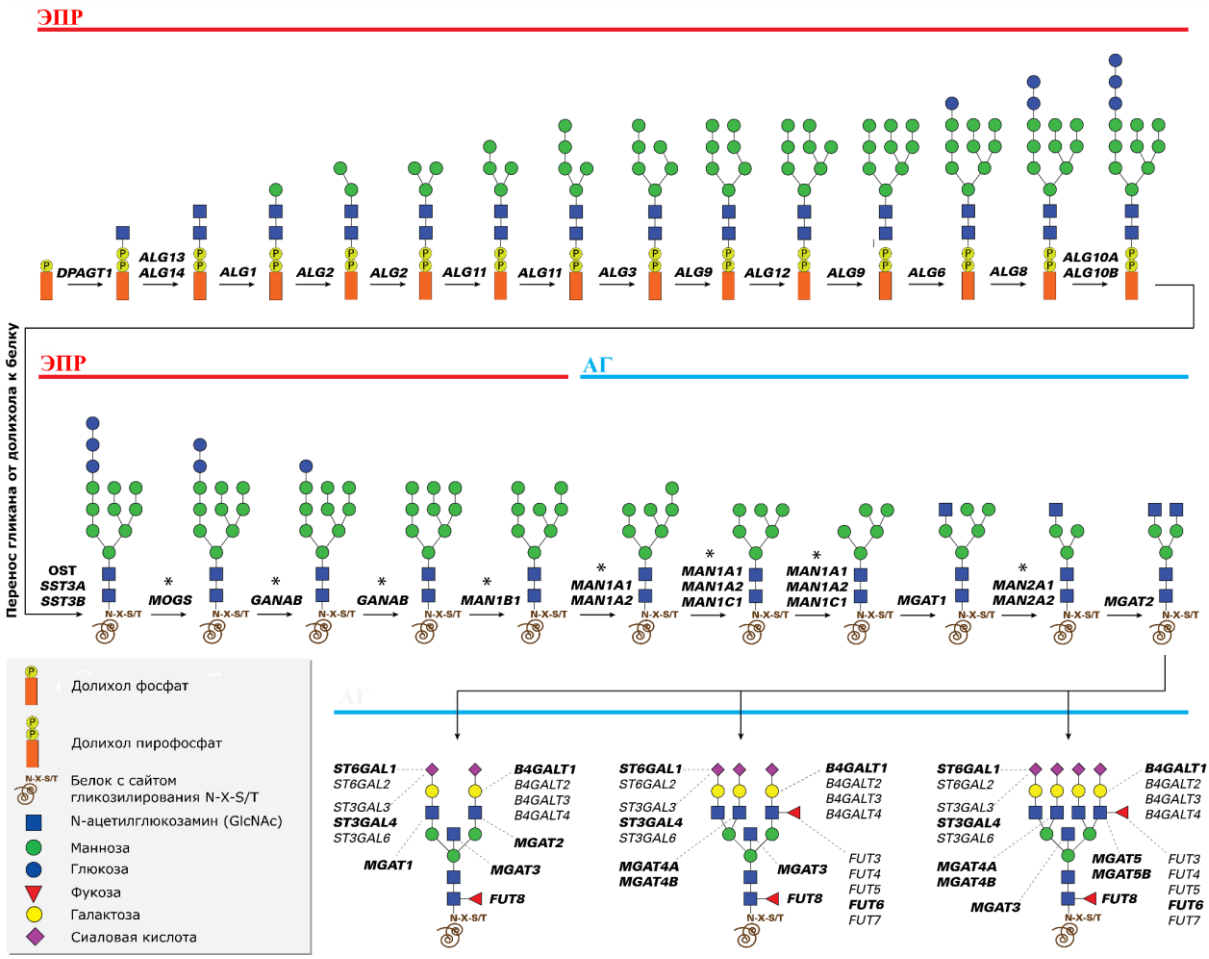


Рис. 3. Биосинтез N-гликанов. Показано разделение биосинтеза N-гликанов на три этапа. Адаптировано из статьи [86].

На Рис. 3 схематически изображен биосинтез N-гликанов. В биосинтезе N-гликанов можно выделить три этапа. В ходе первого этапа в ЭР происходит синтез 14-сахаридного предшественника N-гликанов, связанного с долихолом. Гликозилтрансферазы, кодируемые генами *DPAGT1*, *ALG1-ALG3*, *ALG6*,

ALG8-ALG14, участвуют в синтезе связанного с долихолом предшественника гликана.

В ходе второго этапа происходит перенос предшественника N-гликана с долихола на растущий полипептид ферментом олигосахарилтрансферазой (OST). OST представляет собой белковый комплекс, состоящий из нескольких белков. Гены *STT3A* и *STT3B* кодируют две изоформы каталитической субъединицы белкового комплекса OST. Предшественник гликана переносится к атому азота остатка аспарагина в консервативном мотиве последовательности N-гликозилирования полипептида. После переноса в белок происходит отсечение остатков глюкозы и маннозы от N-связанных гликанов. Гены *MOGS* и *GANAB* кодируют гликозидазы - ферменты, которые удаляют остатки галактозы из гликана, а ген *MAN1B1* кодирует маннозидазу - фермент, удаляющий остатки маннозы из N-гликана. После процессинга в ЭР, белок с N-связанным гликаном переносится из ЭР в АГ, где происходит дальнейшее удаление остатков маннозы маннозидазами (кодируемыми генами *MAN1A1*, *MAN1A2* и *MAN1C1*). В завершении второго этапа происходит образование остова N-гликанов под действием ферментов, кодируемых генами *MGAT1*, *MAN2A1* / *MAN2A2* и *MGAT2*.

В ходе третьего этапа в АГ происходит созревание N-гликанов, в ходе которого остов N-гликанов претерпевает модификации. Ген *FUT8* кодирует фукозилтрансферазу, отвечающую за фукозилирование остова N-гликанов. Ген *MGAT3* кодирует гликозилтрансферазу, необходимую для добавления бисектного N-ацетилглюкозамина (GlcNAc). Образование дополнительных ветвей GlcNAc (т.е. образование третьей и четвертой антенн) катализируется гликозилтрансферазами, кодируемыми генами *MGAT4A*/*MGAT4B* и *MGAT5*/*MGAT5B*. Удлинение ветвей GlcNAc за счет добавления остатков галактозы катализируется гликозилтрансферазами, кодируемыми генами *B4GALT1*/*B4GALT2/B4GALT3/B4GALT*4. Присоединение остатка сиаловой кислоты с образованием альфа 2,6 гликозидной связи происходит с помощью

сиалилтрансфераз, кодируемыми генами *ST6GAL1* и *ST6GAL2*. Присоединение остатка сиаловой кислоты с образованием альфа 2,3 связи катализируется ферментами, кодируемыми генами *ST3GAL3, ST3GAL4* и *ST3GAL6*. Гены *FUT3-7* кодируют фукозилтрансферазы, которые добавляют фукозу к антеннам N-гликанов.

1.4. Методы высокопроизводительного измерения гликанов

Бурное развитие гликобиологии в совокупности с огромным успехом когортных эпидемиологических исследований подтолкнуло разработку методов высокопроизводительного измерения гликома белков плазмы крови. За последние десять лет было разработано несколько методов высокопроизводительного профилирования N-гликома [38] с помощью таких методов, как высокоэффективная и сверхвысокоэффективная жидкостная хроматография (high and ultra-high performance liquid chromatography, HPLC and UHPLC), мультиплексный капиллярный гель-электрофорез с флуоресцентной детекцией (multiplex capillary gel electrophoresis with laser induced fluorescence detection, xCGE-LIF), жидкостная хроматография и тандемная масс-спектрометрия (liquid chromatography electrospray mass spectrometry, LC-MS), и матрично-активированная лазерная десорбция/ионизация и тандемная масс-спектрометр (matrix-assisted laser desorption/ ionization time-of-flight mass spectrometry, MALDI-TOF-MS).

Несмотря на разнообразие технологий, использующихся в данных методах, все они состоят из нескольких ключевых этапов – подготовки образца (клеточной культуры, ткани, органа или организма), выделения гликанов (например, путем отщепления от гликоконъюгатов), разделения гликанов и измерения их концентрации (абсолютной или относительной) [38]. Применение того или иного метода для измерения уровней гликанов имеет свои достоинства и недостатки. Методы MALDI-TOF-MS и LC-MS, основанные на масс-спектрометрии, позволяют проводить сайт-специфичное разделение гликанов, предоставляя тем самым ценную информацию о

гликозилировании конкретных белков. Также данные методы имеют высокую разрешающую способность разделения гликанов с различной молекулярной массой, но при этом они не способны разделить стереоизомеры гликанов. Из-за высокой стоимости, трудозатратности и требований к квалификации сотрудников, число гликомных профилей, измеренных этими методами, невелико.

Наборы гликанных признаков, измеряемые разными методами и даже разными версиями одного и того же метода, отличаются друг от друга, причем варьирует как общее число гликанных признаков, так и образующие их биохимические структуры гликанов. Простого и однозначного соответствия между гликанными признаками, измеряемыми с помощью разных платформ, нет. Таким образом, объединение выборок, образцы N-гликома которых были измерены различными методами, для проведения ПГИА гликома плазмы крови человека на данный момент является невозможным ввиду отсутствия отработанной методологии. Достижение же наибольшей мощности ПГИА гликома плазмы крови требует объединения максимально возможного числа образцов. Среди современных высокопроизводительных методов измерения гликома плазмы крови, наибольшее распространение получил метод СВЭЖХ [39] благодаря его относительной дешевизне, улучшенной (по сравнению с ВЭЖХ) разрешающей способности и высокой производительности. На данный момент с помощью технологии СВЭЖХ были получены профили N-гликозилирования белков плазмы крови для более 20 тысяч образцов.

На рисунках 4 и 5 представлены типичные хроматограммы ВЭЖХ и СВЭЖХ образцов N-гликома плазмы крови. В результате интеграции пиков, для СВЭЖХ обычно выделяют 39 пиков, а на хроматограмме ВЭЖХ – 16 пиков. Для каждого из пиков известен набор гликанных структур, входящих в

состав данного пика. Эти данные получены в результате совмещенного

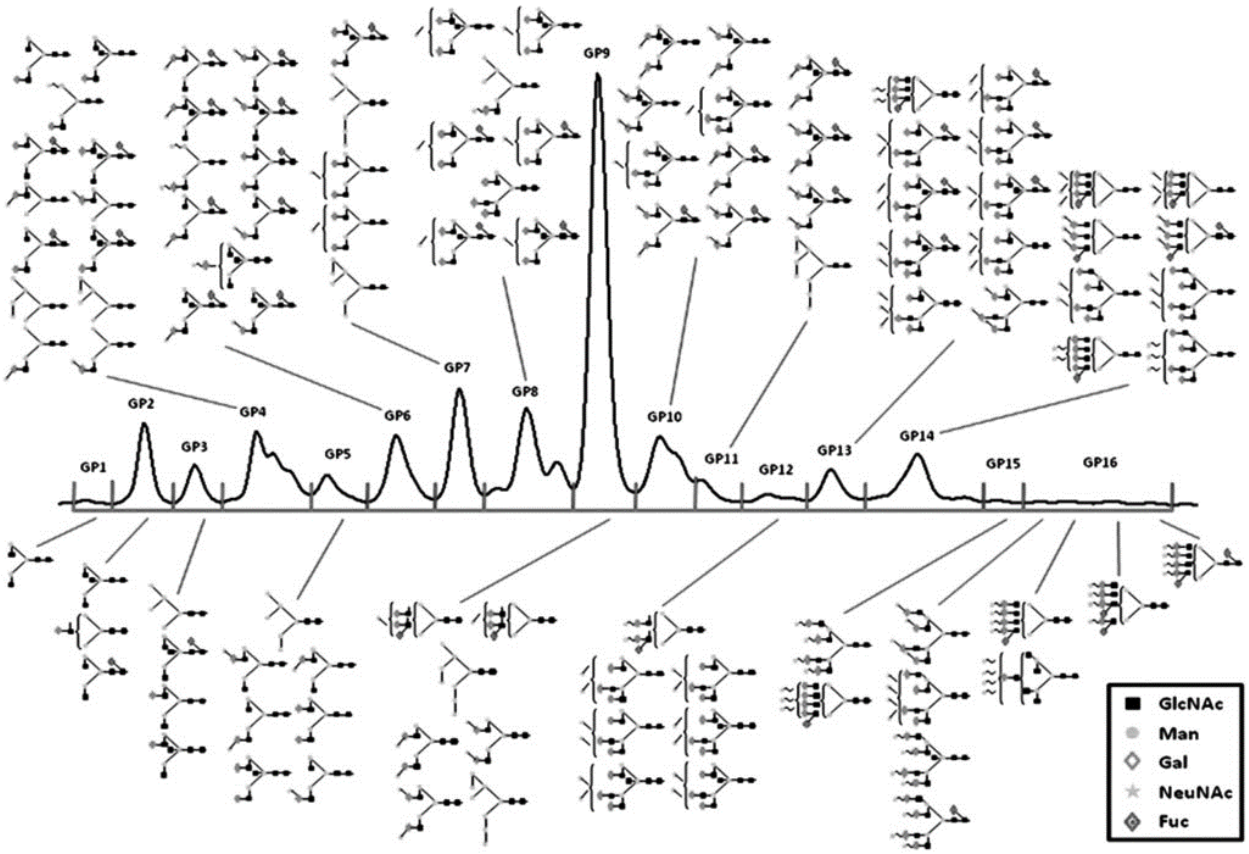


Рис. 4. Пример ВЭЖХ хроматограммы образца N-гликома плазмы крови. По оси X отложено время удерживания, по оси Y – интенсивность сигнала флуоресценции. Каждый пик на хроматограмме аннотирован структурой наиболее представленного N-гликана в данном пике. Взято из [87].

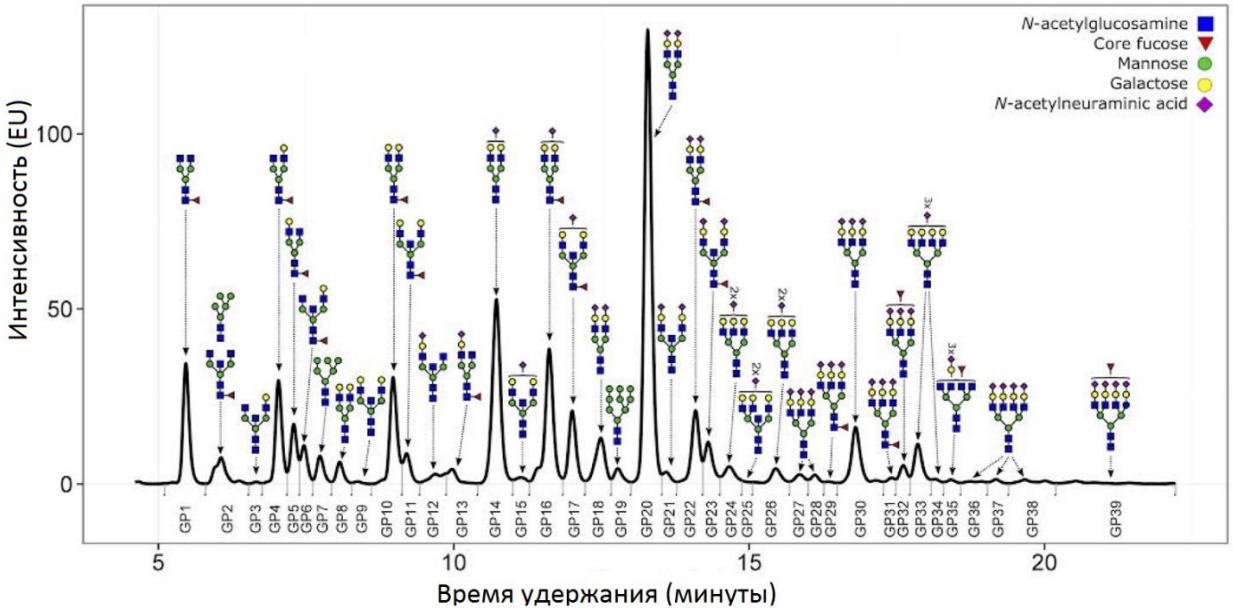


Рис. 5. Пример СВЭЖХ хроматограммы образца N-гликома плазмы крови. По оси X отложено время удерживания, по оси Y – интенсивность сигнала флуоресценции. Каждый пик на хроматограмме аннотирован структурой наиболее представленного N-гликана в данном пике. Взято из [88].

анализа UHPLC-MS, в ходе которого фракция из каждого пика проходит анализ на масс-спектрометре. Видно, что разрешение СВЭЖХ (39 пиков) в разы больше, чем у ВЭЖХ (16 пиков), что позволяет выделить фракции большего числа гликанов с различной структурой.

Одной из методологических трудностей в исследовании N-гликома плазмы крови, измеренного технологией СВЭЖХ на материале нескольких выборок, является изменчивость набора гликанных признаков, уровни которых измерены в разных выборках (см. главу 2.2.5). В зависимости от исследования, число признаков варьирует от 36 до 42 [39, 45–47]. Данные различия обусловлены изменениями в протоколах анализа СВЭЖХ, как на этапах проведения хроматографического анализа (ряд пиков могут оказаться недостаточно разделенными), так и на этапе определения границ пиков – интеграции. Объединение результатов анализов гликома плазмы крови, выполненных на материале нескольких выборок, возможно только при условии того, что в каждой из выборок анализ ассоциаций проводился для единого (гармонизированного) набора признаков. Таким образом, для проведения ПГИА гликома плазмы крови человека с последующей репликацией результатов на независимых выборках, требуется разработка и применение метода гармонизации гликомных профилей СВЭЖХ, полученных в анализируемых выборках.

1.5. Изучение генетического контроля гликозилирования

Развитие методов высокопроизводительного измерения гликома и генетического анализа позволили к началу 2010-х годов провести первые полногеномные исследования генетического контроля гликозилирования на материале когортных исследований. N-гликом плазмы крови стал основным объектом исследования по ряду причин: во-первых, по сравнению с другими тканям человека, плазма крови является более доступным объектом исследования; во-вторых, как было сказано ранее, на данный момент наиболее хорошо разработаны технологии определения структуры и измерения уровней

N-гликанов. Наиболее представленными гликопротеинами плазмы крови человека являются иммуноглобулины G, A, M, фибриноген, трансферрин, гаптоглобин и другие [3]. Главным источником гликопротеинов плазмы крови человека являются клетки печени и клетки, продуцирующие антитела [3, 89].

В работе [90] изучались популяционная изменчивость гликанов плазмы крови человека, их наследуемость (доля дисперсии признака, обусловленная генетическими различиями), а также влияние различных факторов среды на уровни гликанов. Измерения уровней гликанов проводились с использованием технологии ВЭЖХ. В данной работе авторы сделали несколько важных выводов. Во-первых, они обнаружили высокую популяционную изменчивость уровней гликозилирования. Во-вторых, авторы обнаружили достоверный эффект пола и возраста человека на уровень различных гликанов, что говорит о важности учета данных параметров. В-третьих, наследуемость уровней гликанов варьировалась (средний коэффициент наследования h^2= 34,7%, стандартное отклонение - 15.5%), что говорит о том, что гликаны находятся под контролем как генетических, так и средовых факторов. В недавней работе [91] исследователи оценили наследуемость 39-ти N-гликанных признаков, измеренных технологией СВЭЖХ. Было показано, что для 24 из 39 признаков наследуемость составляет более 50% (средний коэффициент наследования h^2=48,0%, стандартное отклонение - 17.7%), что подтверждает гипотезу о существенном влиянии на гликом плазмы крови как средовых, так и генетических факторов.

Не смотря на то, что исследования наследуемости позволяют оценить, какая доля изменчивости признака находится под контролем генома, такие исследования не позволяют выявить конкретные участки генома, влияющие на проявление признака. Поиск таких участков возможен с применением методов картирования генов количественных признаков, в частности, метода полногеномного анализа ассоциаций.

1.6. Полногеномное исследование ассоциаций

Самым широко используемым методом картирования локусов комплексных признаков и заболеваний человека является полногеномное исследование ассоциации (ПГИА). Данный метод предполагает проведение анализа ассоциаций между большим числом (от сотен тысяч до десятков миллионов) генетических маркеров, распределенных по всему геному, и исследуемым признаком. При этом, как правило, анализируются большие (от нескольких тысяч до миллионов) выборки особей или индивидов. Наличие таких данных позволяет тестировать практически весь геном на присутствие ассоциаций с исследуемым признаком и, таким образом, позволяет найти новые, ранее не известные ассоциации между локусами и признаком [92].

Основная идея метода ПГИА заключается в следующем. Аллели близкорасположенных локусов являются сцепленными. Мутация, оказывающая влияние на признак, будет находиться в неравновесии по сцеплению с аллелями близлежащих локусов. Чем ближе расположены локусы, тем дольше они будут наследоваться совместно в ряду поколений. Таким образом, наличие ассоциации между генотипированным маркером и исследуемым признаком позволяет предположить наличие в близлежащих к маркеру геномных районах функционального аллеля, влияющего на значение признака [92].

Анализ ассоциации для количественных признаков проводят с помощью метода линейной регрессии, при этом в уравнении регрессии исследуемый фенотип является зависимой переменной, а генотип – независимой переменной (предиктором). Таким образом, уравнение регрессии можно выписать в виде:

$$E[Y] = \mu + \beta_g * g$$

где Y – фенотип, μ - отступ регрессии, g – генотип ОНП, закодированный как 0, 1 или 2 в зависимости от дозы эффекторного аллеля, β_g

– коэффициент регрессии (в данном случае – размер эффекта эффекторного аллеля на признак). В случае, если исследуемый ОНП не имеет эффект на фенотип, коэффициент регрессии β_g будет незначимо отличен от нуля. Как правило, тестирование отличия коэффициента регрессии проводят с помощью Вальд-теста, тестовая статистика которого имеет вид:

$$T^2 = (b_g/se(b_g))^2$$

где $se(b_g)$ – стандартная ошибка коэффициента регрессии. При нулевой гипотезе $H_0: \beta = 0$ статистика T^2 распределена как χ^2 с одной степенью свободы. В случае, если нулевая гипотеза отвергается, делается вывод о том, что исследуемый генетический маркер имеет эффект на исследуемый признак, либо маркер находится в неравновесии по сцеплению с вариантом, имеющим эффект на исследуемый признак.

При проведении ПГИА, регрессионный анализ проводится для каждого ОНП по отдельности. Общая схема проведения ПГИА представлена на Рис. 6.

После проведения ПГИА оцененные параметры регрессионной модели агрегируют с геномными координатами ОНП, частотой аллеей ОНП в исследованной популяции, метрикой качества импутации [93] и другими характеристиками ОНП в специальный формат данных - суммарные статистики ассоциации [94]. Метрикой качества импутации может являться R^2 - квадрат коэффициента корреляции между вектором восстановленных доз аллелей ОНП, элементы которого принимают континуальное значение на отрезке [0,2] и вектором дискретных генотипов этого же ОНП, элементы которого принимают значения 0, 1 или 2, отражающими лучше предсказание генотипа. Также метрикой качества импутации может быть info score – отношение дисперсии вектора восстановленных доз аллелей ОНП к ожидаемой дисперсии генотипа при условии выполнения равновесия Харди-Вайенберга.

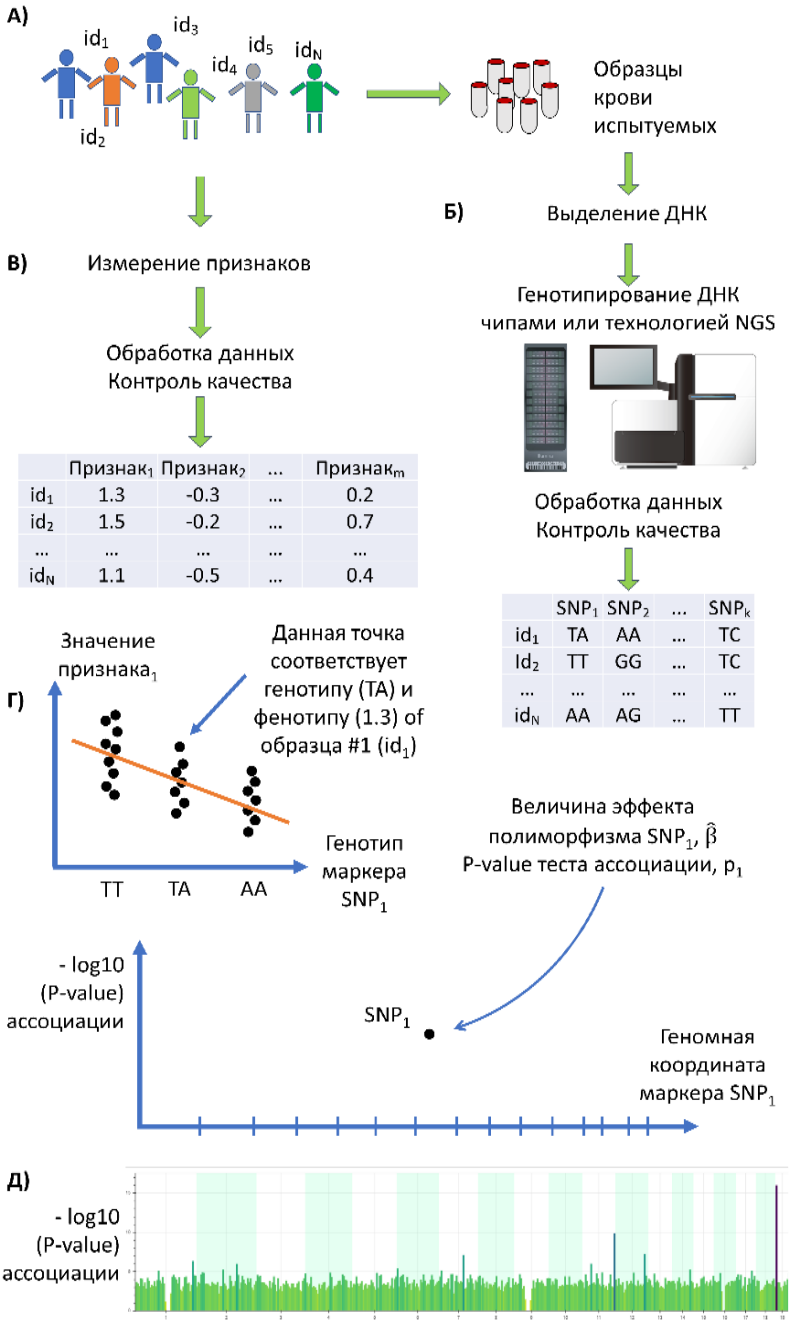


Рис. 6. Общая схема проведения полногеномного анализа ассоциаций на примере количественного признака. Этап А - набор выборки участников исследования и сбор образцов крови. Образцы крови используются для выделения геномной ДНК. Этап Б – выделение ДНК из образцов крови и генотипирование образцов методом ДНК-чипов или ресеквенирования генома с дальнейшим проведением контроля качества геномных данных. Этап В – измерение исследуемого признака и сбор информации о других признаках, важных для проведения исследования (напр. – пол, возраст, статус заболеваний, этническая принадлежность и т.д.). Проведение контроля качества фенотипических признаков. Этап Г – проведение анализа ассоциаций для одного маркера. Анализ проводится методом линейной регрессии, в ходе которого значения исследуемого признака регрессируются против генотипов участников исследования. Этап Д – повторение этапа Г (проведение анализа ассоциаций) для каждого генетического маркера.

Существующая технология, основанная на использовании ДНК-

микрочипов, позволяет генотипировать сотни тысяч маркеров в геномах тысяч особей. Основные принципы генотипирования с помощью ДНК-микрочипов были сформулированы в конце 1980-х годов, однако коммерческие чипы для ПГИА стали доступны для рядового исследователя начиная с 2005-го года. Именно в этом году вышла первая публикация об успешном применении метода ПГИА для исследования сложного признака человека – возрастной дегенерации жёлтого пятна [95]. С выходом этой статьи начался бурный рост числа ПГИА. За последние пятнадцать лет метод ПГИА достиг огромных успехов в картировании локусов, ответственных за признаки и заболевания человека [22, 23]. Например, по состоянию на апрель 2018 года, согласно базе данных GWAS Integrator [96]), с помощью метода ПГИА в 3349 исследованиях найдены полногеномно значимые ассоциации более чем 60 тысяч уникальных ОНП с более чем 14,500 признаками и заболеваниями человека.

Как правило в ПГИА используется материал не одной, а нескольких выборок. Результаты исследования каждой из выборок объединяются с использованием методов полногеномного мета-анализа [49], что увеличивает общий размер выборки, и тем самым, увеличивает статистическую мощность анализа ассоциаций. Мета-анализ позволяет объединять результаты исследований только по тем маркерам, которые присутствуют во всех объединяемых исследованиях. Для того чтобы набор полиморфизмов совпадал между выборками, перед проведением ПГИА проводят «импутацию генотипов» [97] - процесс оценивания апостериорного распределения вероятностей неизмеренных генотипов при условии генотипированных ОНП. Чаще всего импутацию проводят с использованием референтной выборки гаплотипов, полученной в ходе отдельного исследования. В результате импутации геномных данных образцов, измеренных с помощью разных ДНК-чипов, позволяет получить генотипы испытуемых по одинаковому набору полиморфизмов. В настоящее время для импутации генотипов также используются данные о гаплотипах, полученные в ходе проектов «1000 геномов» [98], Haplotype Reference Consortium [41] и TOPMED [42]. Также

существует несколько широко используемых методов импутации генотипов [99]. Среди данных методов нельзя выделить один, превосходящий другие по качеству импутации и производительности [99], поэтому в общем случае выбор метода импутации не принципиально влияет на проведение последующего ПГИА. Импутация данных и мета-анализ позволяют объединить результаты множества независимых исследований, тем самым повышая статистическую мощность картирования локусов и достоверность результатов ПГИА. Это способствует обнаружению ассоциаций признаков с локусами небольшого эффекта, которые на выборках малого объема можно не отличить от случайных ассоциаций. Кроме того, при проведении ПГИА общепринято производить репликацию результатов, т.е. подтверждать найденные ассоциации в дополнительной независимой выборке [100].

Как уже отмечалось ранее, знание локусов является предпосылкой для ответа на вопрос о механизме молекулярно-генетического контроля признака. Ассоциация локуса с признаком объясняется тем, что аллели маркеров в данном локусе находятся в неравновесии по сцеплению с функциональными аллелями, т.е. ассоциация указывает на регион, в котором находится мутация, влияющая на фенотип. В найденных локусах могут располагаться от одного до десятков генов, либо генов в локусе может не быть вообще [22, 23]. При этом существует множество возможных причин возникновения ассоциации – от наличия в локусе кодирующих замен, влияющих на структуру и функционирование продукта гена (белка или РНК) до наличия в локусе замены, влияющей на специфичность связывания транскрипционных факторов с регуляторными областями. Само число функциональных вариантов может варьироваться от одного до многих [101]. Определение функциональных генов в найденных локусах и механизмов их влияния на исследуемый признак является важной задачей функциональных исследований, проводимых с применением методов молекулярной и клеточной биологии. При этом число возможных гипотез, требующих тестирования, растет (теоретически) в геометрической прогрессии в

зависимости от числа возможных молекулярных механизмов ассоциации. Принимая во внимание сложность, дороговизну и трудоемкость методов молекулярной и клеточной биологии, крайне важным является проведение первичной биоинформатической приоритизации гипотез о механизмах возникновения ассоциации. К настоящему времени разработано большое число методов функциональной аннотации *in silico* [52, 55, 56, 94, 101–104], использующих большой спектр накопленных знаний о структуре и функциональной роли замен, генов и участков в геноме [52, 56], экспрессии генов в различных тканях [50, 105, 106], о молекулярно-биологических путях и генных сетях [56]. Применение данных методов позволяет приоритизировать гипотезы о механизме ассоциации, тем самым повышая эффективность дальнейших молекулярно-биологических исследований.

1.7. Полногеномное исследование ассоциаций N-гликома плазмы крови

Общая схема исследования генетического контроля уровней N-гликозилирования белков плазмы крови представлена на Рис. 7. Первые ПГИА N-гликома плазмы крови человека [33, 34] обнаружили шесть локусов, контролирующих уровни N-гликанов. В последнем ПГИА N-гликома плазмы [34] были проанализированы ассоциации 46 гликанных признаков, измеренных технологией ВЭЖХ у 3,533 участников – представителей четырех генетически изолированных выборок из Европы с 2,500,000 ОНП. В результате были обнаружены шесть локусов, ассоциированных с гликомом плазмы на уровне значимости P-value $< 5 * 10^{-8}$ (см. Табл. 1).

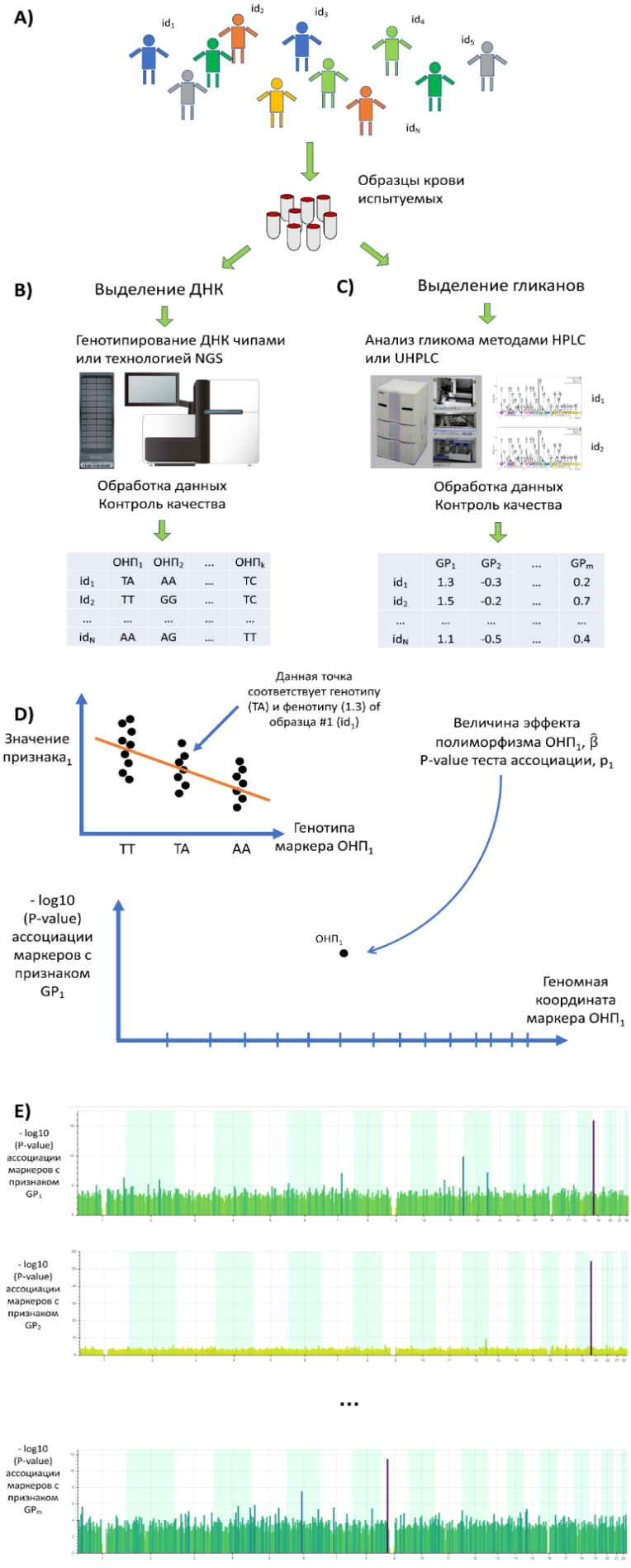


Рис. 7. Обобщенная схема этапов исследования генетического контроля гликозилирования методом ПГИА.

A) Рекрутирование участников исследования и сбор образцов крови. Образцы крови используются для выделения геномной ДНК и получения образцов плазмы крови.

B) Генотипирование образца с использованием ДНК-чипов и/или методов ресеквенирования генома. Результаты генотипирования проходят несколько этапов предобработки и контроля качества, включая импутирование неизмеренных генотипов.

C) Выделение гликанов из образцов плазмы крови и проведение анализа гликома. Результаты измерений проходят несколько этапов предобработки, контроля качества, и последующей обработки, включая вычисление производных признаков.

D) Анализ ассоциации одного генетического маркера с одним гликомным признаком. Анализ ассоциации проводится методом линейной регрессии значений гликомного признака против генотипов маркера. В результате анализа ассоциации оценивается эффект маркера на исследуемый признак и достоверность отличия эффекта от нуля. Отрицательный десятичный логарифм P-value теста ассоциации располагается на оси Y манхэттенского графика, по оси X которого отложена геномная координата маркера ОНП$_1$.

E) Проведение ПГИА анализа ассоциации путем повторения этапа (D) для каждого генетического маркера и каждого гликомного признака.

В трех локусах из шести находились гены, чья роль в биосинтезе N-гликанов была известна ранее [85] – фукозилтрансферазы *FUT6* и *FUT8* и глюкозаминилтрансфераза *MGAT5*. Локус на 11 хромосоме содержит ген, кодирующий фермент *B3GAT1* из семейства глюкуронилтрансфераз, переносящий остаток глюкуроновой кислоты на невосстановленные терминальные углеводы созревающего гликана, в результате чего образуется эпитоп HNK-1. Ранее, присутствие этого эпитопа в составе гликопротеинов плазмы крови не было известно, однако последующий анализ позволит установить его присутствие [34].

Роль двух локусов на хромосомах 3 и 12 (содержащих гены *HNF1A* и *SLC9A9*) в контроле гликозилирования ранее не была известна. Локус на третьей хромосоме содержит ген *SLC9A9*. Этот ген кодирует протонный насос, изменяющий pH в эндосомальном компартменте. Ранее было показано , что изменение pH в аппарате Гольджи влияет на степень сиалирования белков, что может объяснять найденную ассоциацию локуса с гликозилированием белков [107]. Однако, последующее функциональное исследование показало, что ген *HNF1A* (кодирующий фактор транскрипции, экспрессирующийся в органах эндодермального происхождения – печени, почках, поджелудочной железе и т.д.), расположенный на хромосоме 12, регулирует экспрессию генов, кодирующих фукозилтрансферазы (*FUT3, FUT5, FUT6, FUT8, FUT10, FUT11*) в клетках линии HepG2, полученной из ткани печени. Более того, было показано, что *HNF1A* регулирует экспрессию ферментов, необходимых для синтеза ГДФ-фукозы – субстрата фукозилтрансфераз. Таким образом, была показана важная роль гена *HNF1A* в контроле фукозилирования гликанов.

Стоит отметить, что еще в 1996 году было обнаружено [35], что мутации в гене *HNF1A* вызывают сахарный диабет взрослого типа у молодых (Maturity Onset Diabetes of the Young 3, MODY-3) [35, 108]. Основываясь на результатах ПГИА, показавших роль гена *HNF1A* в контроле гликома плазмы крови, были обнаружены потенциальные гликомные биомаркеры заболевания MODY-3

[20]. Эти биомаркеры показали высокую диагностическую точность, что подтверждает важность исследования генетического контроля гликозилирования для разработок методов прогнозирования, диагностирования, профилактики и лечения заболеваний человека.

Табл. 1. Результаты ПГИА гликома плазмы крови, выполненного в работе [34]. ОНП – маркер локуса, позиция – геномные координаты ОНП, маркирующего локус (хромосома: п.н.), GRCh36; Реф/Эфф – референсный/эффекторный аллель; EAF – частота эффекторного аллеля; N – размер выборки, ген – предложенный ген-кандидат в данном локусе, признак – признак с минимальным P-value ассоциации для данного ОНП; BETA / SE – эффект эффекторного аллеля на признак и его стандартная ошибка; P-value - P-value ассоциации ОНП с признаком; Все признаки – все признаки, с которыми ассоциирован данный локус.

ОНП	Позиция	Реф/ Эфф	EAF	N	Ген	Признак	Beta	SE	P-value	Все признаки
rs1257220	2:134731817	A/G	0.7384	3263	*MGAT5*	TA	-0.1860	0.0292	1.80E-10	TA; DG11
rs4839604	3:144442963	A/G	0.7696	3320	*SLC9A9*	Tetra-sialylated	-0.2238	0.0308	3.50E-13	Tetrasialylated
rs7928758	11:133771177	A/C	0.1194	3233	*B3GAT1*	DG13	-0.2285	0.0405	1.66E-08	DG13
rs735396	12:119923227	A/G	0.3949	3236	*HNF1A*	DG7	-0.1848	0.0270	7.81E-12	GP13; GP15; DG7; DG9;DG11;A_FUC
rs11621121	14:64892246	A/G	0.4284	3234	*FUT8*	DG1	0.2674	0.0268	1.69E-23	GP1; GP10; DG1; DG6; DG10; C_FUC;A2;
rs3760776	19:5790746	A/G	0.8699	3262	*FUT6*	DG9	0.4418	0.0394	3.18E-29	GP14; DG7;DG9;DG12;A_FUC

1.8. Краткое заключение

Гликозилирование является важной модификацией белков, которая влияет как на их физико-химические свойства, так и выполняемые ими биологические функции. Гликозилирование белков является тканеспецифичным процессом. Изменения в гликозилировании белков ассоциированы с риском множества заболеваний и признаков человека. С биохимической точки зрения процессы гликозилирования изучены достаточно хорошо, однако регуляция данного процесса *in vivo* остается малоизученной. Применение методов генетического анализа, в частности полногеномного анализа ассоциации, позволяет найти новые, ранее не известные, регуляторы процессов гликозилирования.

Первые полногеномные исследовании ассоциации уровней N-гликанов плазмы крови человека были проведены в 2010-2011 годах с использованием устаревшей на данный момент технологий измерения уровней гликанов (ВЭЖХ) и референтных выборок для импутации (HapMap 3). В течении семи лет, прошедших после публикации последнего ПГИА гликома плазмы в 2011 году, появились новые технологии измерения уровней гликанов, а также новые референтные выборки (например, “1000 Геномов”), разрешение которых на порядок больше таковых, доступных в 2011 году. Использование современной технологии СВЭЖХ для анализа N-гликанов в совокупности с референтными выборками «1000 геномов» позволит обнаружить новые, ранее не известные регуляторы гликозилирования белков плазмы крови человека.

Одной из методологических трудностей проведения ПГИА гликома плазмы крови, измеренного технологией СВЭЖХ, является изменчивость набора гликанных признаков, измеренных в разных выборках. В зависимости от исследования, число признаков варьирует от 36 до 42 [39, 45–47]. Однако, объединение результатов ПГИА нескольких выборок подразумевает, что в каждой из выборок анализ ассоциаций проводился для гармонизированного

набора. Таким образом, для проведения ПГИА гликома плазмы крови человека с последующей репликацией результатов на независимых выборках, требуется разработка и применение метода гармонизации гликомных профилей СВЭЖХ, полученных в анализируемых выборках.

Таким образом, целью данной работы является поиск генов, участвующих в контроле N-гликозилирования белков плазмы крови человека. Для достижения поставленной цели был проведен ПГИА уровней N-гликанов белков плазмы крови человека. Для проверки найденных ассоциаций на материале независимых выборок разработан и валидирован метод гармонизации N-гликомных профилей. Найденные ассоциации были подтверждены на гармонизированном материале независимых выборок. Для найденных локусов проведен биоинформатический анализ для приоритизации генов, в найденных локусах.

Глава 2. Материалы и методы

Данная работа была выполнена на основе материала пяти выборок– TwinsUK, EPIC-Potsdam, PainOmics, SOCCS и SABRE. Демографические, геномные данные и данные измерений уровней N-гликанов белков плазмы крови были любезно предоставлены нашими коллегами в рамках совместного проекта по исследованию генетического контроля гликозилирования белков плазмы крови человека. Описание исследуемых выборок, методики генотипирования и измерения уровней N-гликанов методом СВЭЖХ описаны в разделе «Материалы». Разработанный метод гармонизации гликомных профилей описан в разделе «Разработка и валидация метода гармонизации гликомных профилей». Методы, применявшиеся для проведения контроля качества гликомных и геномных данных, полногеномного анализа ассоциаций, репликативного анализа и биоинформатического анализа локусов приведены в разделе «Методы».

2.1. Схема исследования

Данная работа состоит из четырех этапов, схематически изображенных на Рис. 8. В ходе первого этапа проводилось полногеномное исследование ассоциаций 113 уровней N-гликанов белков плазмы крови человека на материале 2,763 образцов выборки TwinsUK. Для N-гликомных данных был проведен контроль качества, в ходе которого были исключены образцы - статистические выбросы, минимизированы различия в систематической ошибке измерения уровней гликанов. На основе биохимической структуры N-гликанов был предложен набор производных признаков. Для полученных признаков было проведено ПГИА. Далее были определены локусы, достоверно ассоциированные с исследуемыми признаками на уровне

значимости, рассчитанном с учетом коррекции Бонферрони на множественные тестирования.

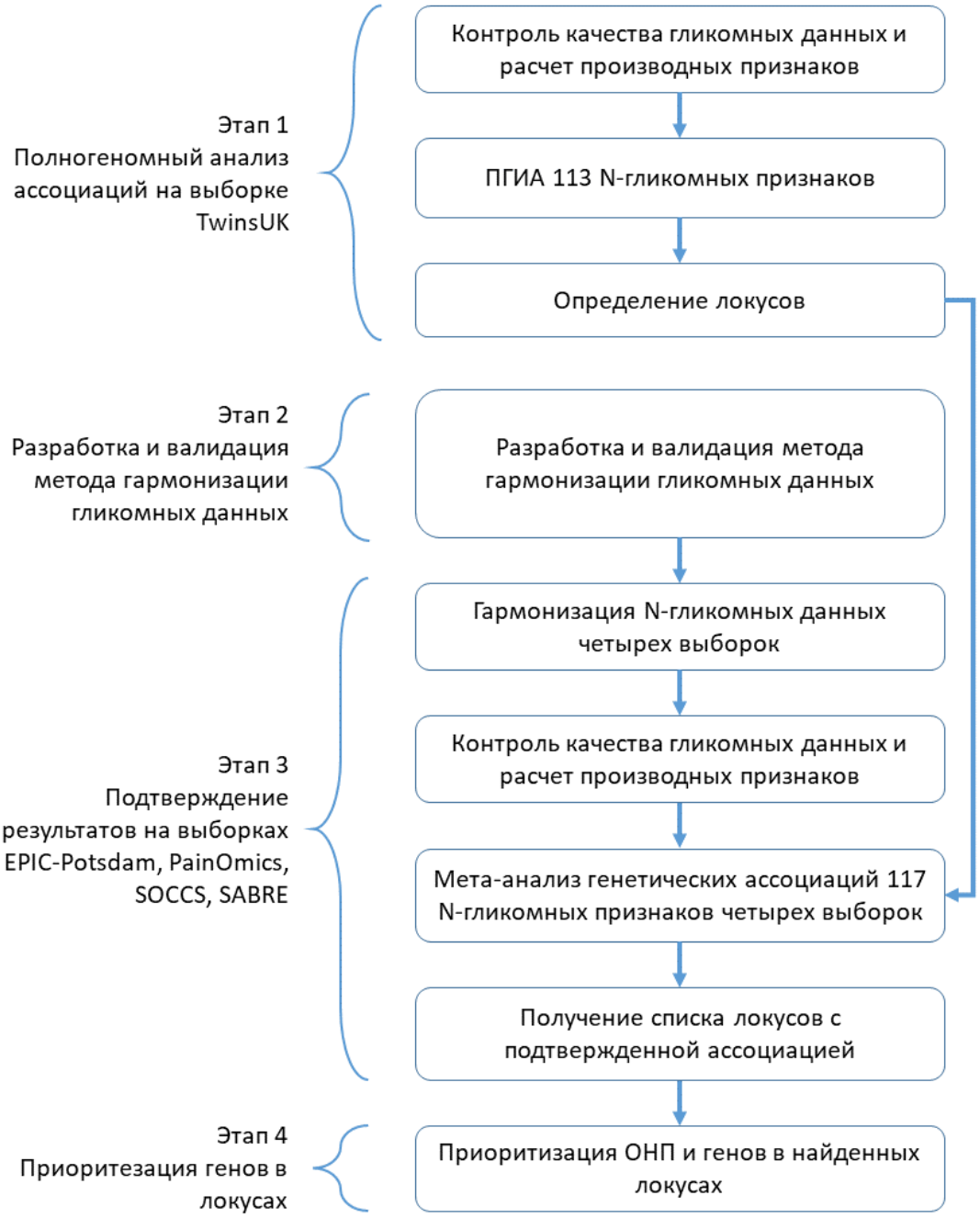


Рис. 8. Схема исследования, выполненного в данной работе.

В ходе второго этапа проводилась разработка и валидация метода гармонизации данных об уровнях N-гликанов белков плазмы крови человека, полученных методом СВЭЖХ для образцов нескольких выборок.

В ходе третьего этапа были подтверждены ассоциации найденных локусов на материале независимых выборок EPIC-Potsdam, PainOmics, SOCCS и SABRE. Уровни N-гликанов 4,802 образцов данных выборок были гармонизированы с использованием метода, разработанного на предыдущем этапе. Для гармонизированных N-гликомных данных был проведен контроль качества. В данном анализе использовалась обновленная (по сравнению с первым этапом) панель гликомных признаков (см. раздел 2.3.4). Объединение результатов анализа генетических ассоциаций найденных локусов, полученных в отдельных выборках, проводилось методом мета-анализа. В результате выполнения третьего этапа была подтверждена ассоциация локусов, найденных на предыдущем этапе.

В ходе четвертого, завершающего, этапа исследования был проведен биоинформатический анализ полученных локусов для приоритизации генов, вероятнее всего вовлеченных в регуляцию N-гликозилирования.

2.2. Материалы

Работа выполнена на материале пяти выборок – TwinsUK, EPIC-Potsdam, PainOmics, SOCCS и SABRE. Описание исследований, в рамках которых были набраны образцы, приведено в Табл. 2. Демографическое описание образцов с измеренным профилем N-гликозилирования белков плазмы крови человека и геномными данными приведено в Табл. 3 (выборки разбиты на подвыборки, см. подробнее раздел 2.3.4, в таблице использованы данные, прошедшие контроль качества). Информированное согласие было получено для всех участников исследования и протоколы исследования были утверждены соответствующими медицинскими этическими комитетами.

Табл. 2. Описание исследований, в рамках которых были набраны выборки людей.

Название выборки	**Название исследование**	**Дизайн исследования**	**Общий размер выборки**	**Этническая группа**	**Ссылка на публикацию**
TwinsUK	TwinsUK Registry	Популяционное	13,000	Британцы	[109, 110]
EPIC-Potsdam	The European Prospective Investigation into Cancer and Nutrition, Potsdam	Популяционное	27,548	Немцы	[111]
PainOmics	PainOmics Retrospective Study	Случай/ контроль. Пациенты с хронической болью в спине и контрольная группа	3,400	Бельгийцы, итальянцы, британцы, хорваты	[88, 112]
SOCCS	Colorectal Cancer Genetics Susceptibility Study	Случай/ контроль. Пациенты с колоректальным раком и контрольная группа	1,762 (1,297 пациента и 465 контрольных образцов)	Шотландцы и другие британцы	[113–115]
SABRE	Southall And Brent REvisited	Популяционное	1,438	Британцы	[116]

Табл. 3. Демографическое описание исследуемых выборок. Сред. возраст – средний возраст; СО возраста – стандартное отклонение возраста.

Название выборки	**Название подвыборки**	**ДНК чип, использовавшийся для генотипирования**	**Число образцов**	**Доля мужчин в %**	**Сред. возраст**	**СО возраста**
TwinsUK	TwinsUK	HumanHap300, HumanHap610Q, 1M-Duo, 1.2MDuo 1M	2,763	10.97	47.83	15.05
EPIC	quad	Human 660W-Quad	325	60.12	50.95	8.88
	corex	Human Core Exome	451	59.73	49.78	8.93
	corex2	Human Core Exome	171	55.81	51.91	8.58
	bonn	Infinium Omni Express Exome	1,245	61.8	50.27	8.93
PainOmics	Belgium	Illumina GSA	149	51.67	52.15	14.24
	UK	Illumina GSA	225	34.67	45.58	14.32
	Italia	Illumina GSA	221	47.96	55.77	15.35
	Croatia	Illumina GSA	156	46.15	47.66	14.69
	Italia	Illumina Human Core	510	47.84	61.20	15.07
	Croatia	Illumina Human Core	190	38.42	52.37	14.19
	Belgium	Illumina Human Core	423	44.79	57.93	13.15
SOCCS	SOCCS controls	Illumina - HumanHap300 и HumanHap240S	459	53.6	51.36	6.29
SABRE	SABRE	Illumina - Human Core Bead Chip	277	90.2	69.71	6.24

2.2.1. Данные исследования TwinsUK

Для проведения ПГИА были использованы данные выборки Британского близнецового исследования – TwinsUK Registry [109, 110]. В рамках данного исследования на добровольной основе проводится анкетирование и сбор биологического материала монозиготных и дизиготных близнецов. В настоящий момент собрана информация о 13,000 близнецов, в основном женщин (83%) среднего возраста. Для 2,763 участников исследования были доступны данные полногеномного генотипирования и первичные данные СВЭЖХ N-гликома белков плазмы крови.

Генотипирование было проведено с использованием геномных чипов компании Illumina - HumanHap300, HumanHap610Q, 1M-Duo и 1.2MDuo 1M. Для данных генотипирования был проведен следующий контроль качества: были исключены ОНП с долей успешно генотипированных образцов < 97% (для ОНП с частотой минорного аллеля ≥ 5%) или < 99% (для ОНП с частотой минорного аллеля < 5%); исключены ОНП с частотой минорного аллеля < 1%; исключены ОНП, не прошедшие тест на равновесие Харди-Вайнберга (P-value $< 10^{-6}$). Образцы с долей прогенотипированных маркеров менее 95% были исключены. Всего, 275,139 ОНП прошли контроль качества.

Импутирование, т.е. восстановление неизмеренных генотипов с помощью информации о гаплотипах образцов референтной выборки [81], было проведено с использованием программного обеспечения (ПО) IMPUTE2 [98] и данных о гаплотипах образцов исследования «1000 геномов» (фаза 1 версия 3, версия сборки генома человека - GRCh37). Для импутированных ОНП был проведен контроль качества – были исключены ОНП с качеством импутирования <70%; с частотой минорного аллеля < 1% и числом копий минорного аллеля < 10. В итоге 8,557,543 ОНП прошли контроль качества и были использованы для проведения ПГИА гликома плазмы крови.

2.2.2. Данные исследования EPIC-Potsdam

Данные исследования EPIC-Potsdam использовались для подтверждения результатов ПГИА N-гликома плазмы крови человека. Проспективное когортное исследование EPIC-Potsdam включает 27,548 участников (16,644 женщины и 10,904 мужчины) в возрастном диапазоне 35-65 лет. Участники исследования набраны случайным образом из популяции в период с 1994 по 1998 годы [111]. Из всех участников, предоставивших кровь в первом исследовании (N = 26,437), была взята случайная выборка участников (N = 2,500), для которых были получены профили N-гликозилирования белков плазмы крови и проведено полногеномное генотипирование. Все участники дали письменное информированное согласие на проведение биомедицинских исследований, и исследование было одобрено Комитетом по этике земли Бранденбург, Германия (Boeing, Korfmann, et al. 1999).

Генотипирование было проведено с использованием геномных чипов компании Illumina: Human660W-Quad (N = 325), HumanCoreExome (N = 622) и InfiniumOmniExpressExome (N = 1245). Для данных генотипирования был проведен контроль качества: исключены ОНП с долей успешно генотипированных образцов <95% (Human660W-Quad, HumanCoreExome) или <96% (InfiniumOmniExpressExome); исключены ОНП, не прошедшие тест на равновесие Харди-Вайнберга (P-value $< 10^{-6}$ для Human660W-Quad и HumanCoreExome и P-value $< 10^{-5}$ для InfiniumOmniExpressExome). Образцы с долей прогенотипированных маркеров менее 95% были исключены. Всего, 564,409 (Human660W) 408,270 (HumanCoreExome), 391,118 (HumanCoreExome) и 849,466 (InfiniumOmniExpressExome) ОНП прошли контроль качества.

Импутирование было проведено с использованием программного обеспечения Eagle [117] и minimac3 [118] и данных о гаплотипах образцов исследования «HRC» версии 1.1 2016 [41], версия сборки генома человека – GRCh37. Для импутированных ОНП был проведен контроль качества – были

исключены ОНП с качеством импутирования <70%; с частотой минорного аллеля < 1% и число копий минорного аллеля < 10. В итоге, следующее число ОНП прошло контроль качества для каждой из подвыборок: 7,338,345 (InfiniumOmniExpressExome, EPIC_bonn), 6,975,419 (HumanCoreExome, EPIC_corex), 5,917,401 (HumanCoreExome, EPIC_corex2), 6,897,446 (Human660W-Quad, EPIC_quad).

2.2.3. Данные исследования PainOmics

Данные исследования PainOmics использовались для подтверждения результатов ПГИА гликома плазмы. PainOmics [112] – исследование типа «случай-контроль», направленное на поиск потенциальных биомаркеров дорсалгии и терапевтических мишеней для его лечения. Письменное информированное согласие было получено от всех участников. В период с сентября 2014 г. по февраль 2016 г. была собрана выборка из 3400 пациентов. Сбор образцов проводился в соответствии со стандартными операционными процедурами, опубликованными в PlosOne в 2017 г. [88].

Генотипирование было проведено с использованием геномных чипов Illumina HumanCore BeadChip и Illumina GSA. Для данных генотипирования был проведен следующий контроль качества: были исключены ОНП с долей успешно генотипированных образцов < 97%; исключены ОНП с частотой минорного аллеля < 1%; исключены ОНП, не прошедшие тест на равновесие Харди-Вайнберга (P-value $< 10^{-5}$). Всего, 301,472 (GSA) и 718,440 (BeadChip) ОНП прошли контроль качества.

Импутирование было проведено с использованием ПО IMPUTE2 [98] и данных о гаплотипах образцов исследования «HRC» версии 1.1 2016 [41], версия сборки генома человека – GRCh37. Для импутированных ОНП был проведен контроль качества – были исключены ОНП с качеством импутирования <70%; с частотой минорного аллеля < 1% и число копий

минорного аллеля < 10. В итоге 3,581,706 (GSA) и 6,843,784 (BeadChip) ОНП прошли контроль качества.

2.2.4. Данные исследования SOCCS

Данные исследования SOCCS использовались для подтверждения результатов ПГИА гликома плазмы. Шотландское исследование SOCCS [119, 120] является исследованием «случай-контроль», направленное на изучение факторов риска колоректального рака. В рамках исследования были собраны данные о 2,057 пациентов с колоректальным раком (61% мужчин; возраст пациентов 65.8 ± 8.4) и 2,111 контрольных испытуемых (60% мужчин; средний возраст 67.9±9.0). Данные исследования SOCCS включают в себя 472 участников контрольной группы, у которых был измерен гликом плазмы крови человека методом СВЭЖХ.

Генотипирование было проведено с использованием комбинации геномных чипов компании Illumina - HumanHap300 и HumanHap240S. Для данных генотипирования был проведен следующий контроль качества: были исключены ОНП с долей успешно генотипированных образцов < 95%; с частотой минорного аллеля < 1%; не прошедшие тест на равновесие Харди-Вайнберга (P-value $< 10^{-6}$). Всего, 514,177 ОНП прошли контроль качества.

Импутирование было проведено с использованием ПО SHAPEIT [121] и ПО IMPUTE2 [98] и данных о гаплотипах образцов исследования «1000 геномов» (фаза 1 версия 3, версия сборки генома человека - GRCh37). Для импутированных ОНП был проведен контроль качества – были исключены ОНП с качеством импутирования <70%; с частотой минорного аллеля < 1% и число копий минорного аллеля < 10. В итоге 7,381,694 ОНП прошли контроль качества.

2.2.5. Данные исследования SABRE

SABRE - это популяционное исследование, начатое в 1988 году [116]. Всего были собраны данные о 4,857 участников исследования в возрасте от 40

до 69 лет, проживающих в Западном Лондоне, Великобритания. Анализ N-гликозилирования белков плазмы крови человека был проведен с использованием биологических образцов, собранных при повторном посещении участников исследования в 2008-2011 гг. Все участники дали письменное информированное согласие.

Генотипирование 1,429 участников исследования было проведено с использованием геномных чипов компании Illumina - HumanCoreBeadChip. Для данных генотипирования был проведен следующий контроль качества: были исключены ОНП с долей успешно генотипированных образцов < 95% исключены ОНП с частотой минорного аллеля < 1%; исключены ОНП, не прошедшие тест на равновесие Харди-Вайнберга (P-value $< 10^{-4}$). Всего, 332,849 ОНП прошли контроль качества.

Импутирование было проведено с использованием программного обеспечения Eagle [117] и minimac3 [118] и данных о гаплотипах образцов исследования «HRC» версии 1.1 2016 [41], версия сборки генома человека - GRCh37. Для импутированных ОНП был проведен контроль качества – были исключены ОНП с качеством импутирования <70%; с частотой минорного аллеля < 1% и число копий минорного аллеля < 10. В итоге 6,593,822 ОНП прошли контроль качества.

2.2.6. Данные гликома плазмы крови, измеренные технологией СВЭЖХ

Анализ профиля N-гликозилирования белков плазмы крови в исследуемых выборках (за исключением выборки SOCCS) проводился в лаборатории «Genos», г. Загреб, Хорватия. Детальный протокол эксперимента опубликован и доступен в статье Трбоевич-Акмачич и коллег [39]. Вкратце, N-гликаны были выделены из плазмы крови в результате обработки образцов ферментом PNGase F, отщепляющий N-гликаны от белков. В отщепленные N-гликаны вводилась флуоресцентная метка (2-аминобензамид, 2-AB), присоединяющаяся к остову N-гликана. Несвязанные молекулы 2-AB

удалялись из образцов методом HILIC-SPE (HILIC - hydrophilic interaction liquid chromatography - жидкостная хроматография за счет гидрофильного взаимодействия, SPE – твердофазная экстракция). Гликаны, меченные флуоресцентной меткой, разделялись прибором Acquity UHPLC («Waters», США). Измерение уровней гликанов проводилось с помощью флуоресцентного детектора FLR с длинами волн возбуждения и эмиссии 250 и 428 нм соответственно. Система измерения была калибрована с использованием внешнего стандарта – 2-AB меченых олигомеров глюкозы известной концентрации.

Анализ профиля N-гликозилирования белков плазмы крови для образцов SOCCS был выполнен в National Institute for Bioprocessing Research and Training (NIBRT, Национальный институт исследований и обучения биотехнологии, Дублин, Ирландия) с применением протокола, описанного выше, с единственной разницей в длине волны возбуждения (330 нм вместо 250 нм).

Интеграция (определение границ пиков на полученных хроматограммах) результатов анализа N-гликома образцов выборок TwinsUK, SABRE, SOCCS и PainOmics (только подвыборка из Италии) проводилась в ручном режиме. Интеграция результатов хроматографии образцов выборок EPIC-Potsdam и PainOmics (все подвыборки, кроме итальянской) проводилась с использованием автоматического метода [37]. Далее, каждая из хроматограмм была скорректирована в ручном режиме для получения одинаковых интервалов интегрирования для всех образцов.

В зависимости от выборки, число интервалов для интегрирования варьировалось от 36 до 42. Варьирование числа пиков связано с несколькими причинами. Во-первых, СВЭЖХ образцов выборок TwinsUK, EPIC Potsdam, PainOmics (Рис. 10) и SABRE (Рис. 9) проводилось в научном центре Genos по одному протоколу, а измерение профилей образцов выборки SOCCS проводилось в центре NIBRT согласно другому протоколу (см. раздел

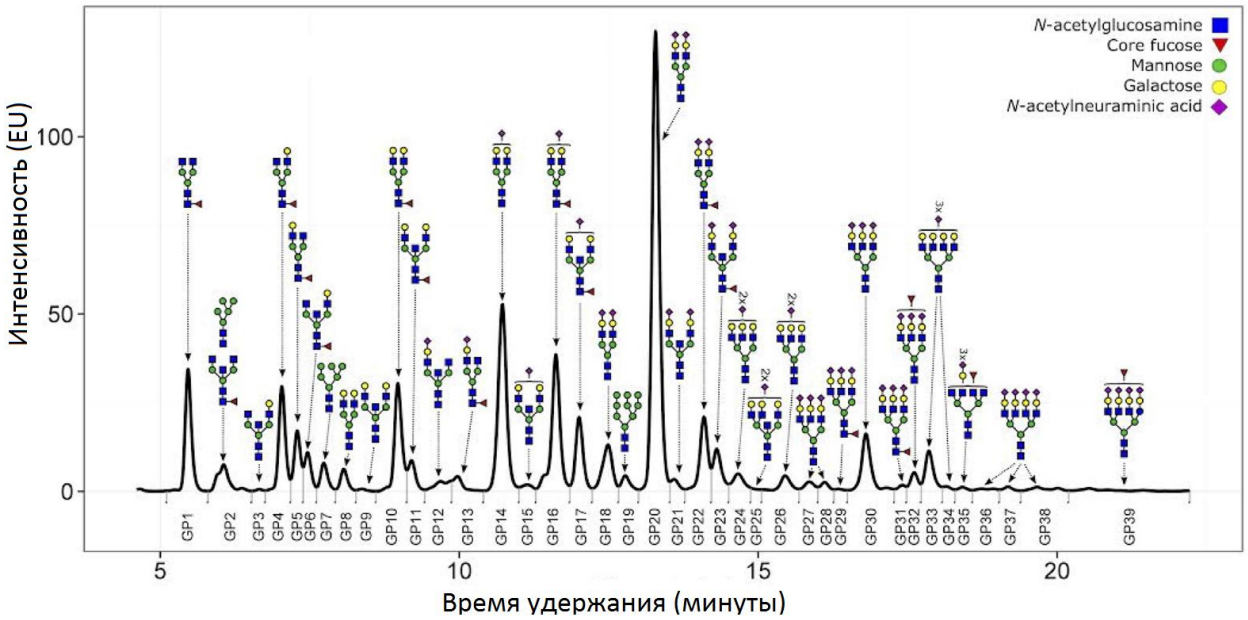


Рис. 9. Пример СВЭЖХ хроматограммы с 36 пиками, полученной в выборке SABRE [122]. По оси X отложено время удерживания, по оси Y – интенсивность сигнала флуоресценции. Каждому образу приписаны структуры гликанов, входящих в состав пика согласно аннотации, использовавшейся до 2019 года [47].

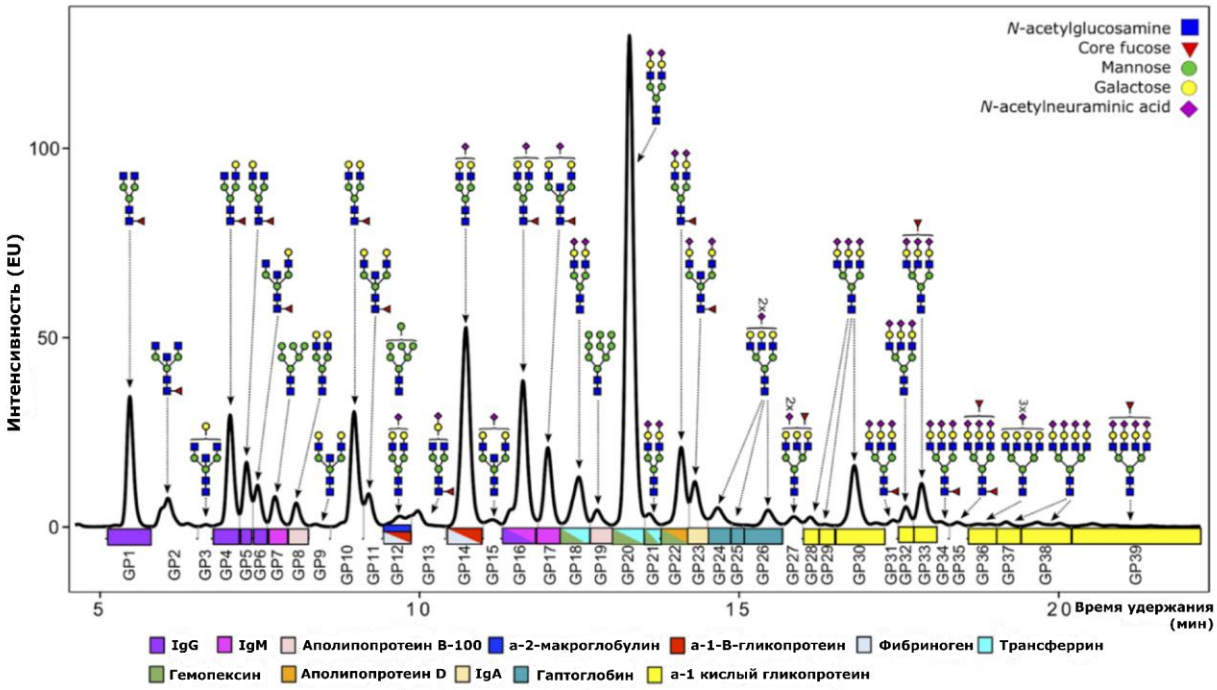


Рис. 10. Пример СВЭЖХ хроматограммы, полученной для образа из выборки TwinsUK. По оси X отложено время удерживания, по оси Y – интенсивность сигнала флуоресценции. Каждому образцу приписаны структуры гликанов, входящих в состав пика согласно аннотации 2020 года [91]

Материалы). В результате N-гликомные профили хроматограмм образцов выборки SOCCS (Рис. 11) имеют большее разрешение - по сравнению с выборкой SABRE шесть пиков были разделены на два. Во-вторых, при интеграции N-гликомных профилей образцов выборки SABRE

(Рис. 9) не были разделены три пика по сравнению с образцами выборок TwinsUK, EPIC-Potsdam и PainOmics (Рис. 10). При этом вне зависимости от разрешения СВЭЖХ, порядок выхода компонентов (N-гликанов) из колонки не изменяется.

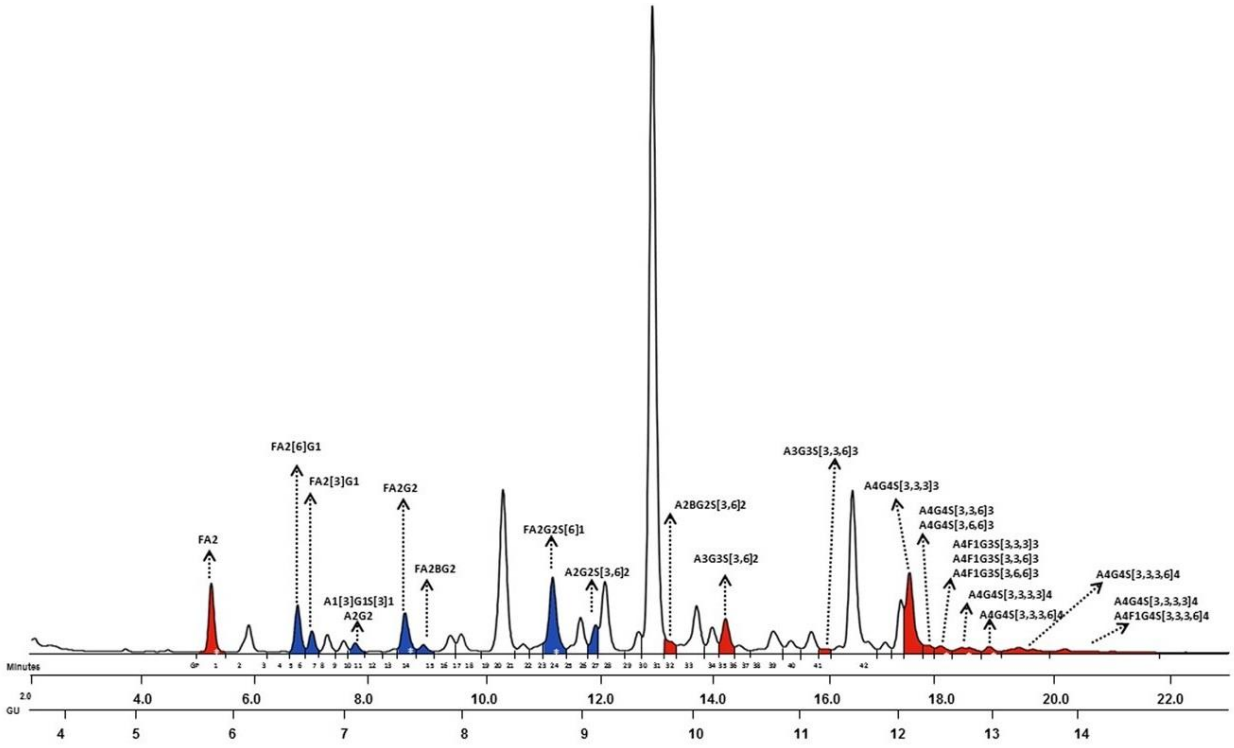


Рис. 11. Пример СВЭЖХ хроматограммы с 42 пиками, полученной для образа из выборки SOCCS. По оси X отложено время удерживания, по оси Y – интенсивность сигнала флуоресценции.

Площадь под каждым из пиков является количественным признаком, отражающим уровень N-гликанов с известной биохимической структурой, формирующих данный пик. Как правило, один из N-гликанов является доминирующим, то есть вносит основной вклад в формировании пика на хроматограмме, и его структура используется в интерпретации результатов анализа данного пика. На Рис. 10 приведена аннотация пиков, использовавшаяся до 2019 года [47]. В 2020 году была опубликована новая аннотация гликомных пиков структурами N-гликанов [91]. Данная аннотация приведена на Рис. 11. В новой аннотации для восьми пиков была обновлена мажорная структура N-гликана, при этом шесть из восьми пиков расположены в правой части хроматограммы, то есть имеют более долгое время удерживания.

Таким образом, для проведения ПГИА гликома плазмы крови человека на выборке TwinsUK с последующим подтверждением результатов на независимых выборках (EPIC Potsdam, PainOmics, SOCCS и SABRE) было необходимо гармонизировать гликомные профили, полученные в анализируемых выборках. Для этого в данной работе мы разработали протокол гармонизации гликомных профилей (см. раздел 3.2 Разработка и валидация метода гармонизации гликомных профилей) и применили его для гармонизации гликомных данных исследуемых выборок. В результате гликомные данные каждой из выборок содержали значения 36 гликомных пиков.

2.3. Методы

2.3.1. Контроль качества гликомных данных

Контроль качества является неотъемлемой частью анализа гликомных данных. Основными целями контроля качества являются устранение образцов - статистических выбросов и коррекция сдвига систематической ошибки измерения в разных сериях. Контроль качества гликомных признаков для выборки TwinsUK проводился следующим образом:

1. Нормализация на общую площадь. Для каждого образца площадь каждого из пиков делится на общую площадь всех пиков данного образца и умножается на 100;

2. Логарифмирование нормализованной площади каждого из пиков;

3. Удаление образцов – статистических выбросов (образец является статистическим выбросом, если значение хотя бы одного из пиков выходит за три межквартильных отклонения от его среднего значения);

4. Устранение сдвига систематической ошибки измерения в разных сериях методом ComBat [123]. Данный метод использует информацию о номере эксперимента и положении образца в рамках эксперимента (например,

номер 96-ти луночного планшета и положение образца на планшете) для минимизации различий в средних и дисперсии измерений между экспериментами, в результате чего распределение систематической ошибки приводится к гомогенному;

5. Повторное удаление статистических выбросов (как и в п.3);
6. Расчет производных признаков на основе биохимической структуры N-гликанов в 36 пиках. Перед расчетом производится экспоненцирование признаков. Расчет производных признаков проводился согласно аннотации гликомных пиков версии 2019 года (см. Рис. 10).

Данный протокол контроля качества был применен для гликомных данных образцов выборки TwinsUK, использовавшихся для проведения полногеномного анализа ассоциаций, результаты которого были опубликованы в 2019 году [124].

В 2019-2020 годах были опубликованы результаты двух исследований, показавших, что протокол контроля качества гликомных данных может быть улучшен. Работа [125], показала, что для нормализации гликомных профилей оптимальным может являться метод Probabilistic Quotient Normalization – метод вероятностных квантилей, а не принятый до этого метод нормализации на общую площадь. В исследовании [91] была опубликована новая аннотация N-гликомных профилей, измеряемых технологией СВЭЖХ, в которой для восьми пиков была предложена новая структура мажорного N-гликана (Рис. 10). Поэтому, в следующем исследовании, целью которого являлось подтверждение результатов анализа ассоциации на материале независимых выборок (EPIC Potsdam, PainOmics, SOCCS и SABRE), протокол контроля качестве гликомных данных был обновлен. Вместо нормализации на общую площадь, использовался метод вероятностных квантилей, а для расчёта производных признаков была использована аннотация профиля N-гликанов белков плазмы крови 2020 года.

2.3.2. Полногеномное исследование ассоциаций

Полногеномное исследование ассоциаций было проведено на данных 2,763 участников исследования TwinsUK для 113 гликанных признаков и 8,557,543 ОНП. Все признаки были предварительно скорректированы на эффекты пола и возраста и приведены к нормальному распределению методом квантильной нормализации, поскольку генетический анализ генетических ассоциации с использованием линейной аддитивной модели предполагает нормальное распределение изучаемого признака [126]. Нарушение данного предположения может привести к увеличению ошибки I рода [126]. Для проведение ПГИА использовалось ПО «GEMMA» [127]. Использовалась смешанная линейная модель ассоциации, учитывающая возможную генетическую структурированность исследуемой выборки. Для анализа импутированных данных использовалась регрессия на оцененное число аллелей. Полученные тестовые статистики был скорректированы методом геномного контроля [128] для устранения возможной инфляции тестовой статистики анализа ассоциаций.

Метод геномного контроля основан на выявлении инфляции значений тестовых статистик, полученных в результате проведения ПГИА, по отношению к ожидаемому распределению тестовых статистик при нулевой гипотезе. Инфляция тестовой статистики может говорить о наличии популяционной стратификации в исследуемой выборке, либо о некорректности выбора метода для ее учета, либо одновременно и о том и том.

Как отмечалось в разделе 1.6, при нулевой гипотезе статистика T^2 теста ассоциации распределена как χ^2 с одной степенью свободы. Одной из часто используемых метрик инфляции тестовой статистики является λ_{median}, определенная как отношение медианы наблюдаемых тестовых статистик всех генетических маркеров к медиане распределения $\chi 2$ с одной степенью свободы, которая равна 0.455 [128]. При предположении о том, что доля

маркеров, ассоциированных с признаком, невелика, медиана распределения тестовой статистики не должна существенно смещаться из-за наличия сигналов ассоциации. В случае отсутствия популяционной стратификации и сигналов ассоциации ожидается, что $\lambda_{median} = 1$. В случае отсутствия популяционной стратификации и наличии маркеров, ассоциированных с признаком, λ_{median} будет незначительно больше 1 – обычно в пределах 1.00-1.05. В данном случае, λ_{median} можно использовать для проведения геномного контроля, а именно для коррекции значения тестовой статистики простым делением тестовой статистики T^2 каждого из маркеров на λ_{median}. В результате данной процедуры медиана наблюдаемых тестовых статистик всех генетических маркеров станет равна 0.455 и фактор инфляции примет значение 1.

2.3.3. Определение локусов

В данном исследовании ПГИА проводилось для каждого из 113 признаков. Поскольку общее число проведенных статистических тестов было велико, для сохранения ошибки первого рода на уровне 5% требовалось выбрать соответствующий уровень значимости. Для идентификации ОНП, достоверно ассоциированных с гликомом, был выбран уровень значимости (с учетом коррекции Бонферони на множественное тестирование) $5 * 10^{-8} / 30 = 1.66 * 10^{-9}$, где $5 * 10^{-8}$ – общепринятый полногеномный уровень статистической значимости, а 30 – число главных компонент, суммарно объясняющих более 99% вариации N-гликома белков плазмы крови образцов выборки TwinsUK. Сигнал ассоциации считался значимым, если значение P-value теста ассоциации не превышало уровень $1.66 * 10^{-9}$.

ОНП, показавшие значимую ассоциацию, были объединены в локусы согласно правилу: ОНП находятся в одном локусе, если они расположены на одной хромосоме и расстояние между ними не превышает 500,000 п.н. Из всех ОНП, находящихся в одном локусе, выбирался ОНП с наименьшим P-value ассоциации (т.е. наиболее достоверный) для данного признака. В случае, если

в локусе находились несколько ОНП, достоверно ассоциированные с разными признаками, то выбирался признак и ОНП с наименьшим P-value ассоциации.

Для обозначения локуса в тексте использовались геномные координаты (номер хромосомы и расположение ОНП на хромосоме с точностью до м.п.н.) ОНП, показавшего наименьшее P-value ассоциации в данном локусе. Например, для обозначения локуса, маркером которого является ОНП rs1169303, расположенный на 12 хромосоме, 121,436,376 п.н., используется выражение: «локус на 12 хромосоме, 121 м.п.н.».

2.3.4. Подтверждение результатов ПГИА на независимых выборках

Для подтверждения ассоциации локусов, были использованы данные четырех выборок - EPIC-Potsdam (N=2,192), PainOmics (N=1,874), SOCCS (N=459) и SABRE (N=277). Суммарный объем репликационный выборки составил 4,802. Анализ генетических ассоциаций найденных локусов проводился на данных выборках с использованием того же протокола, что и для поискового ПГИА, выполненного на выборке TwinsUK. Объединение результатов, полученных на четырех выборках, проводилось мета-анализом с помощью метода обратных дисперсий при предположении фиксированных эффектов. Мета-анализ проводился с пользованием ПО METAL [129]. В данном анализе использовалась обновленная (по сравнению с данными выборки TwinsUK) панель гликомных признаков. Поэтому тестировалась ассоциация найденных локусов со всеми 117 N-гликомными признаками из обновленной панели. Использовался уровень значимости P-value $< 0.05/(16 * 117) = 2.67 * 10^{-5}$, где 16 – число локусов, а 117 – число N-гликомных признаков. Ассоциация локуса считалась подтвержденной, если P-value ассоциации было меньше использованного порога.

2.3.5. Оценка мощности анализа ассоциаций на независимых выборках

Оценка мощности анализа ассоциаций проводилась следующим образом. При альтернативной гипотезе (H1 : $beta \neq 0$) тестовая статистика ассоциации $T^2 = \left(\frac{beta}{se}\right)^2$, оцененная на выборке распределена как $\chi^2_{\mathrm{df}=1,\mathrm{NCP}}$, где NCP – параметр нецентральности, оцениваемый по формуле

$$NCP = \left(T^2_{disc} - 1\right) * {N_{rep}}/{N_{disc}}$$

где $T^2_{disc} = \left(\frac{beta_{disc}}{se_{disc}}\right)^2$ – тестовая статистика ассоциации ОНП с признаком, оцененная на поисковой выборке, N_{rep} – объем независимой выборки, использующейся для подтверждения найденных ассоциаций, и N_{disc} – объем выборки, на которой был проведен поиск ассоциаций. Мощность анализа ассоциаций равна вероятности того, что тестовая статистика превысит порог значимости.

2.3.6. Определение доверительного набора ОНП и их функциональная аннотация

Общепринятым форматом представления результатов ПГИА является составление сводной таблицы ассоциации найденных локусов. В данной таблице каждый локус представлен генетическим маркером, показавшим наименьшее (наиболее достоверное) P-value ассоциации в данном локусе. Однако ассоциация данного маркера далеко не всегда является каузальной [130]. С одной стороны, это обусловлено тем, что не все генетические варианты представлены в ДНК-чипе. Даже процесс импутации с использованием большого числа гаплотипов референтной выборки не гарантирует того, что каузальные замены будут измерены. С другой стороны, даже если каузальные варианты измерены в исследуемых выборках, есть шанс того, что их ассоциация будет не самой достоверной в локусе [131].

Для решения задачи тонкого картирования – определения доверительного набора ОНП, замены в которых наиболее вероятно являются каузальными – разработано множество методов. В данной работе был составлен доверительный набор ОНП на основе эвристического подхода [130]. В доверительный набор вошли ОНП, удовлетворяющие хотя бы одному из двух критериев:

1. ОНП находится в неравновесии по сцеплению (r^2>0.6) с наилучшим ОНП, показавшим самое низкое P-value ассоциации в данном локусе.
2. ОНП, расположен в +/-250 т.п.н. от наилучшего ОНП. При этом для данного ОНП P-value<T, где T определялся как $\log_{10}(T)=\log_{10}(Pmin)+1$, где Pmin - P-value наилучшего ОНП в локусе.

Неравновесие по сцеплению рассчитывалось в подгруппе образцов европейского происхождения (503 человека) выборки 1000 Genomes Project phase 3 version 5. Второй критерий был использован, поскольку геномные данные образцов выборки TwinsUK, на материале которой проводился ПГИА, были импутированы с использованием данных о гаплотипах образцов исследования «1000 геномов» (фаза 1 версия 3, версия сборки генома человека - GRCh37). Поэтому не все восстановленные ОНП присутствовали в геномных данных, использовавшихся для оценки неравновесия по сцеплению.

Для ОНП из полученного набора была проведена функциональная аннотация. Были определены ОНП, замены в которых приводят к изменению первичной последовательности белков. Для этого использовался программный пакет VEP (Variant Effect Predictor) [52]. Данный программный пакет использует широкий набор данных об экспрессии генов в тканях, регуляторных районах, частот встречаемости замен в популяциях человека для аннотации и приоритизации геномных вариантов в кодирующих и не кодирующих областях, а также для оценки влияния замены (ОНП,

инсерции/делеции) на гены, транскрипты, белки или транскрипционные факторы. В случае, если замена в ОНП из доверительного набора расположена в кодирующей последовательности гена и приводит к изменению первичной последовательности белка, данный ген включался в набор генов-кандидатов.

Также были предсказаны возможные последствия нуклеотидных замен с помощью методов FATHMM-XF [53] и FATHMM-INDEL [54]. Эти программные пакеты используют метод скрытых марковских моделей и ядерные методы машинного обучения, классифицирующие замены на нейтральные и патогенные. Данный алгоритм был обучен на выборке патогенных замен, собранных в базе Human Gene Mutation Database [132] и выборке нейтральных замен из данных «1000 геномов».

2.3.7. Анализ биологических путей и тканеспецифичной экспрессии

Для идентификации биологических путей и тканей, в которые вероятнее всего вовлечены гены в найденных локусах, использовался метод DEPICT [56]. Данный метод основан на предположении о том, что гены, ассоциированные с исследуемым признаком, должны иметь сходную функциональную аннотацию – например, быть вовлечены в один биологический путь или экспрессироваться в одних и тех же тканях.

Метод DEPICT использует 14.461 реконструированных наборов генов (“gene sets”). Данные наборы генов были составлены на основе аннотации генов, извлеченных из различных источников, включая реконструированные экспертами молекулярные пути (manually curated pathways); молекулярные пути, реконструированные из сетей белок-белковых взаимодействий; наборы генов со схожим фенотипическим проявлением после нокаута в мышах. В методе DEPICT заложены рассчитанные вероятности принадлежности каждого из генов каждому их реконструированных наборов генов. DEPICT оценивает уровень перепредставленности каждого из наборов генов среди всех генов, расположенных в предложенных ему локусах (например, локусах,

ассоциированных с исследуемым признаком). В случае, если какой-то из наборов генов показал достоверный уровень обогащения в исследуемых локусах, делается вывод о возможной роли данного набора генов в формировании исследуемого признака, а также приоритизируются гены, входящие в данный набор.

Помимо этого, метод DEPICT использует данные о экспрессии генов в 209 тканях человека для приоритизации тканей и клеточных типов, в которых высоко экспрессируются гены в исследуемых локусах.

2.3.8. Анализ плейотропных эффектов на экспрессию генов

Более 90% генетических ассоциаций, найденных методом ПГИА, расположены в некодирующих областях генома [130]. Данные области обогащены регуляторными элементами (промоторами, энхансерами и т.д.). Также известно, что ОНП, найденные методом ПГИА, более вероятно ассоциированы с уровнем экспрессии близлежащих генов, чем другие ОНП, генотипируемые ДНК-чипами. Здесь и далее под уровнем экспрессии гена понимается количественная характеристика числа молекул мРНК, синтезированных по матрице последовательности гена. Другими словами, локусы, найденные ПГИА, чаще других проявляют плейотропные эффекты на уровень экспрессии близлежащих генов. Таким образом, считается, что подавляющая часть сигналов ассоциаций, определяемая ПГИА, обусловлена влиянием замен на экспрессию генов, которая в свою очередь, приводит к изменению признака.

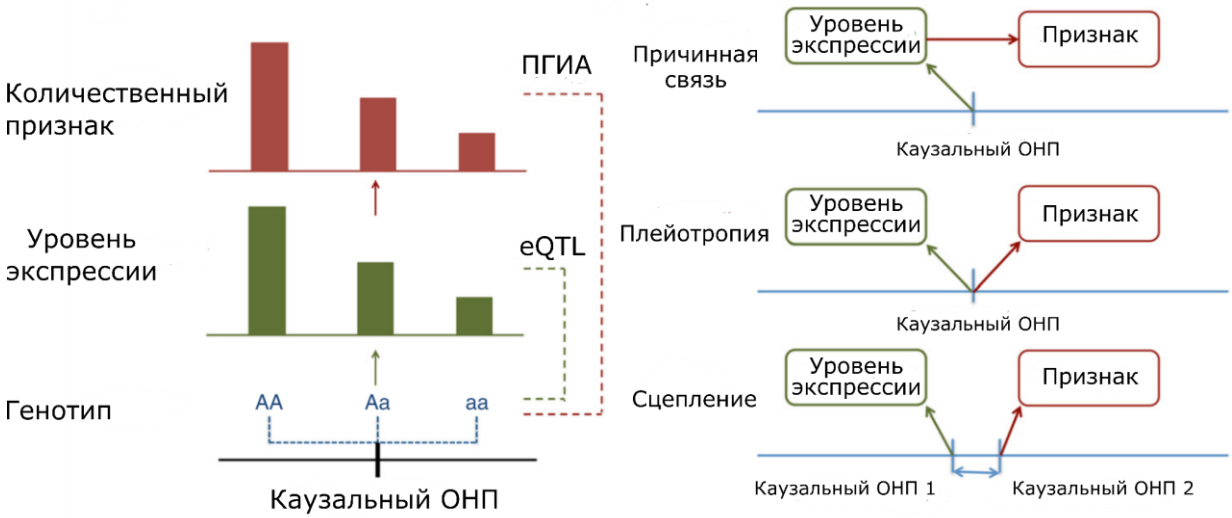
Для определения генов, чей уровень экспрессии может опосредовать найденные ассоциации между локусом и признаков, используются результаты анализа cis-eQTL эффектов. Если ОНП в найденном локусе ассоциированы с уровнем экспрессии определенного гена, то возможны три сценария взаимосвязи между ОНП, экспрессией гена и признаком (см. Рис. 12). В сценарии «причиной связи», ОНП в локусе имеет плейотропные эффекты на

уровень экспрессии гена и исследуемый признак, при этом изменение уровня экспрессии влияет на исследуемый признак. В сценарии «плейотропии» ОНП в локусе имеет плейотропные эффекты на уровень экспрессии гена и исследуемый признак, однако экспрессия гена и исследуемый признак не связаны между собой. В сценарии «сцепления» в локусе присутствуют сцепленные ОНП, независимо влияющие на признак и на уровень экспрессии гена.

Для определения генов, чей уровень экспрессии контролируется найденными сигналами ассоциации, применялся метод Менделевской рандомизации (SMR) с последующим анализом гетерогенности (HEIDI).

Метод SMR тестирует ассоциацию между уровнем экспрессии гена и признаком. Пусть x – уровень экспрессии определенного гена (или его изоформы), y – значение исследуемого признака и z – генетический маркер (ОНП). $\hat{b}_{zy}$ – оценка эффекта ОНП на признак и $\hat{b}_{zx}$– оценка эффекта ОНП на уровень экспрессии гена, полученные в результате соответствующих ПГИА. Тогда оценка эффекта уровня экспрессии гена на исследуемый признак описывается формулой $\hat{b}_{SMR} = \frac{\hat{b}_{zy}}{\hat{b}_{zx}}$, где $\hat{b}_{SMR}$– оценка эффект уровня экспрессии гена на исследуемый признак. Для тестирования эффекта $\hat{b}_{SMR}$ используется статистика: $T_{SMR} = \frac{\hat{b}^2_{SMR}}{var(\hat{b}_{SMR})} \approx \frac{(z^2_{zy} z^2_{zx})}{(z^2_{zy} + z^2_{zx})}$, где $z_{zy} = \hat{b}_{zy}/se_{zy}$; $z_{zx} = \hat{b}_{zx}/se_{zx}$; se_{zy} и se_{zx}– стандартные ошибки оценки эффектов ОНП на исследуемый признак и уровень экспрессии гена соответственно. Нулевая гипотеза SMR теста – ассоциация между уровнем экспрессии гена и признаком отсутствует, альтернативная гипотеза – уровень экспрессии гена ассоциирован с признаком. В случае, если нулевая гипотеза отвергается, возможны три сценария, описанные на Рис. 12.

Рис. 12. Ассоциация между локусом и количественным признаком, опосредованная экспрессией гена в локусе. Слева на рисунке: причинно-следственная модель, в которой изменение в признаке опосредовано изменением в уровне экспрессии определенного гена. Справа на рисунке: возможные механизмы обнаруженной ассоциации между локусом, транскрипцией гена и признаков. Причинная связь – каузальный ОНП изменяет экспрессию гена, вследствие чего изменяется фенотип. Плейотропия – функциональный ОНП влияет независимо на уровень экспрессии и на признак. Сцепление – в локусе находятся два функциональных варианта в неравновесии, влияющих на уровень экспрессии гена и на признак. Рисунок адаптирован из [55].



HEIDI тест позволяет разграничить ситуации «причинной связи» или «плейотропии» от «сцепления». HEIDI использует информацию о направлении и величине эффекта ОНП, находящихся в локусе, на уровень экспрессии гена и на признак. Статистика HEIDI теста определена следующим образом: $T_{HEIDI} = \sum_i^m z_{d(i)}^2$, где m – число ОНП в локусе, выбранных для анализа, $z_{d(i)} = {d_i}/{SE_{(d_i)}}$ и $d_i = \hat{b}_{SMR_i} - \hat{b}_{SMR\ (lead\ SNP)}$, где lead SNP – ОНП с минимальным P-value ассоциации с исследуемым признаком. При нулевой гипотезе теста HEIDI распределение эффектов ОНП в локусе («паттерн» ассоциации) на уровень экспрессии гена и на признак совпадают. В этом случае делают вывод о том, в локусе может находиться функциональный вариант (или варианты), плейотропно влияющий на уровень экспрессии гена и на признак.

При альтернативной гипотезе, распределение эффектов ОНП в локусе на уровень экспрессии гена и на признак не совпадают, таким образом, делают вывод о том, в локусе находится два разных функциональных варианта,

влияющих на уровень экспрессии и на признак независимо. Более подробное математическое описание методов SMR/HEIDI доступно в работе [55]. Следует, однако, понимать, что два аллеля, каждый из которых оказывает влияние на свой признак (верна гипотеза «сцепления»), не могут быть различены тестом HEIDI, если они находятся в идеальном неравновесии по сцеплению.

Для проведения SMR/HEIDI анализа были использованы публично доступные данные консорциума GTEx [50] и Westra [105]. Данные GTEx включает в себя информацию о цис- и транс-эффектах 11,552,519 ОНП на уровни экспрессии генов, измеренные постмортально в 44 тканях технологией RNA-Seq у 450 людей. Данные Westra содержат информацию о цис- и транс-эффектах 1,962,237 ОНП на уровни экспрессии 29,891 изоформ генов, измеренных в периферической крови у 5,311 людей. Дополнительно в рамках сотрудничестве с Университетом Льежа (г. Льеж, Бельгия) нам были предоставлены данные выборки CEDAR, в которых содержится информация о цис- и транс-эффектах ОНП на уровни экспрессии генов, измеренные в 9 клеточных линиях и тканях, включая CD4+ (Т-хэлперы), CD8+ (Т-киллеры), CD14+ (макрофаги), CD15+ (лакунарные гистиоциты), CD19+ (В-лимфоциты), CD45- (тромбоциты).

Граничный уровень значимости SMR теста был установлен как $P\text{-value}_{SMR} < 0.05/20{,}448 = 2.45 * 10^{-6}$ (где 20,448 – общее число различных изоформ генов (с учетом всех тканей), использовавшихся в анализе. Для HEIDI теста использовались следующие уровни значимости – $P_{HEIDI} > 0.05$ – общий генетический контроль уровня экспрессии гена и гликома плазмы вероятен, $0.001 < P_{HEIDI} < 0.05$ – общий генетический контроль уровня экспрессии гена и гликома плазмы возможен, $P_{HEIDI} < 0.001$ – общий генетический контроль уровня экспрессии гена и гликома плазмы отсутствует или мало вероятен.

2.3.9. Определение потенциальных плейотропных эффектов локусов на комплексные признаки и заболевания человека

Найденные локусы, ассоциированные с N-гликозилированием, были проверены на наличие ассоциаций с другими признаками. Для этого использовалась база данных “PhenoScanner”, которая содержит в себе информацию о более чем 3 миллиардах ассоциаций для более чем 10 миллионов уникальных ОНП [104]. Из базы данных была извлечена информация об ассоциациях набора ОНП, полученного в разделе 2.3.6, с комплексными признаками и заболеваниями человека. Для ассоциаций был выбран порог значимости P-value $< 5 * 10^{-8}$.

Глава 3. Результаты

3.1. ПГИА и определение локусов, ассоциированных с уровнями N-гликанов белков плазмы крови человека

Ранее были опубликованы два полногеномных исследований ассоциаций (ПГИА) уровней N-гликанов белков плазмы крови человека [33, 34], целью которых являлось выявление новых генов-регуляторов N-гликозилирования белков. В этих исследованиях уровни N-гликанов были измерены с помощью технологии ВЭЖХ. В пилотном исследовании [33] ПГИА было проведено для 13 признаков на материале трех выборок (N=2,705). В результате были найдены три локуса - *HNF1A*, *FUT6*/*FUT3* и *FUT8*. В 2011 году в исследовании [34] к трем выборкам была добавлена четвертая (суммарный N=3,533). Число гликомных признаков было увеличено до 46 (33 непосредственно измеренных и 13 производных признаков) за счет проведения ВЭЖХ двух образцов N-гликанов плазмы крови испытуемых: 1) необработанный образец (16 признаков) 2) образец после обработки десиалилирующим агентом (13 признаков). Дополнительно была проведена ВЭЖХ со слабым анионообменным носителем (4 признака). Геномные данные были импутированы до 2,500,000 ОНП с использованием референтной выборки HapMap [133]. В результате были найдены три новых локуса – *B3GAT1*, *MGAT5* и *SLC9A9*.

В данной работе впервые проведен ПГИА уровней N-гликанов белков плазмы крови, измеренных технологией СВЭЖХ, имеющей лучшие (по сравнение с ВЭЖХ) точность и разрешение. Были проанализированы геномные данные, импутированные с использованием современной референтной выборки «1000 геномов» [134]. Эти данные полногеномного ресеквенирования предоставляют лучшее (по сравнению с HapMap) покрытие генома и точность восстановления генотипов.

Полногеномное исследование ассоциаций с целью выявления новых локусов, ассоциированных с уровнями N-гликанов белков плазмы крови человека, было проведено на выборке TwinsUK. Перед проведением ПГИА, гликомные и геномные данные образцов выборки TwinsUK прошли контроль качества. На основе полученных полногеномных суммарных статистик ассоциаций было проведено определение локусов, показавших ассоциацию на выбранном уровне значимости P-value $< 1.7 * 10^{-9}$. В завершении было проведено сравнение полученных результатов с опубликованными ранее.

3.1.1. Контроль качества N-гликомных данных и расчет производных признаков

Данные N-гликома белков плазмы крови 2,816 участников исследования TwinsUK прошли контроль качества, описанный в разделе «Методы». В результате 53 образца (1.9% от общего числа) были исключены, а 2,763 образца успешно прошли контроль качества.

Далее было проведено вычисление производных признаков. Производные признаки являются линейной комбинацией или отношением двух линейных комбинаций признаков и вычисляются на основе сходства биохимической структуры N-гликанов, входящих в состав 36 пиков на хроматограмме. Использование производных признаков является общепринятой практикой в гликомных исследованиях в целом, и в ПГИА гликома в частности. Считается, что включение производных признаков в анализ позволяет увеличить мощность анализа ассоциации [33, 34, 45, 64]. В рамках данной работы нами был составлен список производных признаков для данных СВЭЖХ плазмы крови. Гликаны были разбиты на подгруппы согласно следующим характеристикам:

1. Наличие или отсутствие рассечения остова;
2. Ветвление сахарной цепи - ди/три/тетра антеннарные гликаны;
3. Наличие или отсутствие фукозилирования N-ацетилглюкозамина в составе остова;

4. Наличие или отсутствие фукозилирования антеннарных цепей;
5. Галактозилирование антеннарных цепей: отсутствие галактозилирования /моно /ди /три /тетра галактозилирование;
6. Сиалирование антеннарных цепей: нейтральные гликаны /моно /ди /три /тетра сиалирование.

На основе данной классификации гликанов были сформулированы производные признаки, например, «доля нейтральных гликанов», «доля моносиалилированных гликанов», «доля дигалактозилированных гликанов с рассечением остова» и т.п. Формирование признаков было осуществлено по подобию существующего списка производных признаков для СВЭЖХ данных N-гликозилирования IgG [64].

Для выборки TwinsUK список производных признаков и формулы их расчета были составлены на основе аннотации профиля СВЭЖХ N-гликома плазмы крови, принятой до 2019 года [47]. В итоге были рассчитаны 77 производных признаков, и общее число признаков для образцов выборки TwinsUK составило 113. Описание производных признаков и формулы расчета приведены в Доп. табл. 2.

3.1.2. Полногеномное исследование ассоциаций

Перед проведением ПГИА все признаки были предварительно скорректированы на эффекты пола и возраста и приведены к нормальному распределению методом квантильной нормализации.

ПГИА был проведен для 113 N-гликомных признаков и 8,557,543 ОНП, измеренных у 2,763 участников исследования TwinsUK. Использовалась смешанная линейная модель ассоциации, учитывающая возможную генетическую структурированность выборки. Для анализа импутированных данных использовалась регрессия на оценённые дозы аллелей. Таким образом,

в данном исследовании впервые проведен ПГИА уровней N-гликанов белков плазмы крови человека, измеренных технологией СВЭЖХ.

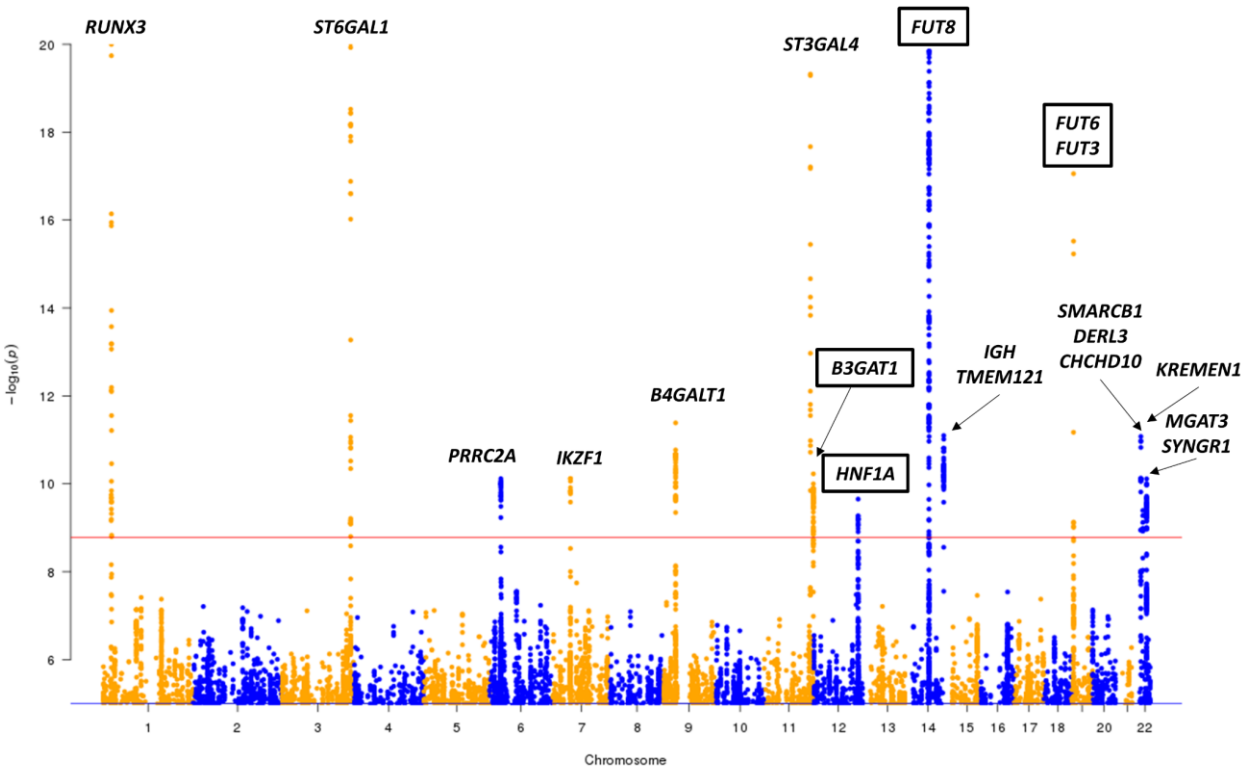


Рис. 13. Результаты ПГИА гликома плазмы крови человека. Показано расположение локусов, ассоциированных с гликомом плазмы крови человека. По оси X - геномная координата ОНП (номерами обозначены номера хромосом). По оси Y – отрицательный десятичный логарифм P-value ассоциации: $-\log_{10}$(P-value). Для каждого ОНП выбрано наименьшее P-value среди 113 признаков. Красная линия соответствует уровню значимости ($1.7 * 10^{-9}$). На рисунке показаны ОНП с P-value $< 1 * 10^{-5}$. Точки с $-\log_{10}$(P-value)>20 показаны на уровне $-\log_{10}$(P-value)=20. Локусы, прошедшие порог отмечены названиями генов, приоритизированных в результате биоинформатического анализа (см. Раздел 3.4). Рамкой выделены локусы, найденные в предыдущих исследованиях.

В общей сложности, 908 ОНП показали значимую (P-value $< 1.7 * 10^{-9}$) ассоциацию как минимум с одним из 68 гликанными признаком (всего обнаружено 5,090 значимых ассоциаций ОНП-признак, см. Рис. 13. Для разных признаков, коэффициент геномного контроля λ варьировался от 0.99 до 1.02 (см. Доп. Табл. 4), что говорит о слабом влиянии популяционной стратификации на полученные оценки ассоциаций. Полученные статистики ассоциаций для признаков с $\lambda > 1$ были скорректированы на коэффициент геномного контроля. В результате ассоциация 906 ОНП осталась

полногеномно значимой. Значимо ассоциированные ОНП были расположены в 14 локусах (см Табл. 4 и Рис. 13).

Мы сравнили полученные результаты с опубликованными ранее [33, 34]. Для четырех локусов из четырнадцати ассоциация была показана ранее. Ассоциация трех локусов – 12 хромосома 121 м.п.н. (ОНП rs1169303, расположенный в интроне гена *HNF1A*), 14 хромосома 66 м.п.н. (ОНП rs7147636, расположенный в интроне гена *FUT8*) и 19 хромосома 58 м.п.н (ОНП rs7255720, расположен в 3’ некодируемой области гена *FUT3*) была показана в обеих работах [33, 34], в то время как ассоциация локуса на 11 хромосоме в 126 м.п.н. (ОНП rs1866767, расположенный в интроне гена *B3GAT1*) была описана только в работе [34].

Табл. 4. Результаты ПГИА гликома плазмы крови человека. В верхней части таблицы приведены результаты ПГИА десяти новых локусов; в нижней - результаты ПГИА для четырех локусов, ассоциация которых с гликомом была показана ранее. Позиция – хромосома и позиция ОНП в геномных координатах (сборки генома GRCh37); ген – предложенный кандидатный ген в локусе; Эфф/Реф – эффекторный/референтный аллель; EAF – частота эффекторного аллеля; BETA (SE) – оценка эффекта эффекторного аллеля на признак и его стандартная ошибка; P – P-value ассоциации ОНП с признаков; P_{GC} – P-value ассоциации после коррекции геномным контролем; признак – признак с минимальным P-value ассоциации для данного ОНП; число признаков – число признаков, ассоциированных с ОНП на выбранном уровне значимости.

Локус				**TwinsUK**					
ОНП	Позиция	Ген	Эфф / Реф	EAF	BETA (SE)	P	P_{GC}	Признак	Число признаков
rs186127900	1:25318225	*RUNX3*	G/T	0.99	-1.26 (0.119)	1.35E-24	4.04E-24	FBG1n/G1n (PGP82)	26
rs59111563	3:186722848	*ST6GAL1*	D/I	0.74	0.34 (0.031)	9.51E-27	1.09E-26	FG1S1/(FG1+FG1S1) (PGP41)	3
rs3115663	6:31601843	*HLA*	T/C	0.80	0.26 (0.040)	6.31E-11	7.65E-11	M9 (PGP18)	1
rs6421315	7:50355207	*IKZF1*	G/C	0.59	0.19 (0.029)	7.50E-11	7.57E-11	A2[6]BG1n (PGP60)	2
rs13297246	9:33128617	*B4GALT1*	G/A	0.83	-0.26 (0.038)	3.46E-12	4.11E-12	FA2G2n (PGP67)	2
rs3967200	11:126232385	*ST3GAL4*	C/T	0.88	-0.49 (0.043)	8.64E-28	1.51E-27	A2G2S[3,6+3]2 (PGP17)	7
rs35590487	14:105989599	*IGH; TMEM121*	C/T	0.77	-0.24 (0.034)	5.90E-12	7.98E-12	FA2[3]G1n (PGP62)	2
rs9624334	22:24166256	*SMARCB1: DERL3; CHCHD10*	G/C	0.85	0.28 (0.040)	4.79E-12	8.38E-12	FA2[6]BG1n (PGP63)	2
rs140053014	22:29550678	*KREMEN1*	I/D	0.98	-0.67 (0.106)	4.01E-10	4.05E-10	G3S2/G3S3 (PGP109)	1
rs909674	22:39859169	*MGAT3*	C/A	0.27	0.22 (0.033)	6.31E-11	7.72E-11	FBS2/FS2 (PGP56)	3
rs1866767	11:134274763	*B3GAT1*	C/T	0.87	0.28 (0.043)	5.15E-11	5.95E-11	A4G4S[3,3,3,3]4 (PGP33)	3
rs1169303	12:121436376	*HNF1A*	A/C	0.51	0.19 (0.029)	2.07E-10	2.23E-10	A4G4S[3,3,3]3 (PGP30)	2
rs7147636	14:66011184	*FUT8*	T/C	0.33	-0.39 (0.030)	4.71E-37	6.63E-37	FA2G2S[3+6,6+3]2 (PGP20)	17
rs7255720	19:5828064	*FUT3; FUT6*	G/C	0.96	1.14 (0.068)	2.13E-55	2.53E-55	G4S3/G4S4 (PGP110)	18

3.1.3. Краткое заключение

В данном исследовании впервые проведено ПГИА уровней представленности N-гликанов белков плазмы крови человека, измеренных технологией СВЭЖХ. Получены суммарные статистики ассоциаций для 113 N-гликомных признаков и 8,557,543 ОНП. Сравнительный анализ полученных результатов с опубликованными ранее показал, что среди найденных локусов присутствуют четырех (из шести локусов), выявленных ранее в анализе выборок, фенотипированных методом ВЭЖХ. Мы впервые показали ассоциацию десяти локусов генома с уровнями N-гликанов белков плазмы крови. Таким образом, общее число локусов, с когда-либо показанной ассоциацией с уровнями N-гликанов белков плазмы крови человека было увеличено с 6 до 16. Полученные результаты опубликованы в работе [124]

3.2. Разработка и валидация метода гармонизации гликомных профилей

Данное исследование проводилось на материале нескольких выборок. На материале выборки TwinsUK проводилось ПГИА и определение локусов. Выборки EPIC Potsdam, PainOmics, SOCCS и SABRE использовались для подтверждения ассоциаций найденных локусов. Для этого на материале каждой из выборок был проведен анализ генетических ассоциаций 16 локусов с уровнями N-гликанов белков плазмы крови человека. Далее результаты, полученные в отдельных выборках, были объединены методом мета-анализа. Объединение результатов анализа ассоциаций нескольких выборок подразумевает, что в каждой из выборок анализ ассоциаций проводился для одного и того же (гармонизированного) набора признаков. Однако гликомные профили плазмы крови человека, полученные технологией СВЭЖХ, могут содержать от 36 до 42 пиков на хроматограмме [39, 45]. Поэтому нами был разработан метод гармонизации гликомных профилей, измеренных технологией СВЭЖХ. Данный метод был валидирован на образцах выборок

TwinsUK и PainOmics. Была показана эффективность данного метода для проведения гармонизации N-гликомных профилей.

3.2.1. Разработка метода

N-гликомные профили исследуемых выборок содержали от 36 до 42 пиков, измеренных методом СВЭЖХ. В зависимости от эксперимента некоторые из пиков могут оказаться неразделимыми на хроматограмме (в результате чего они объединяются в один пик), при этом порядок пиков и структуры N-гликанов, входящие в эти пики, сохраняются. Данная ситуация

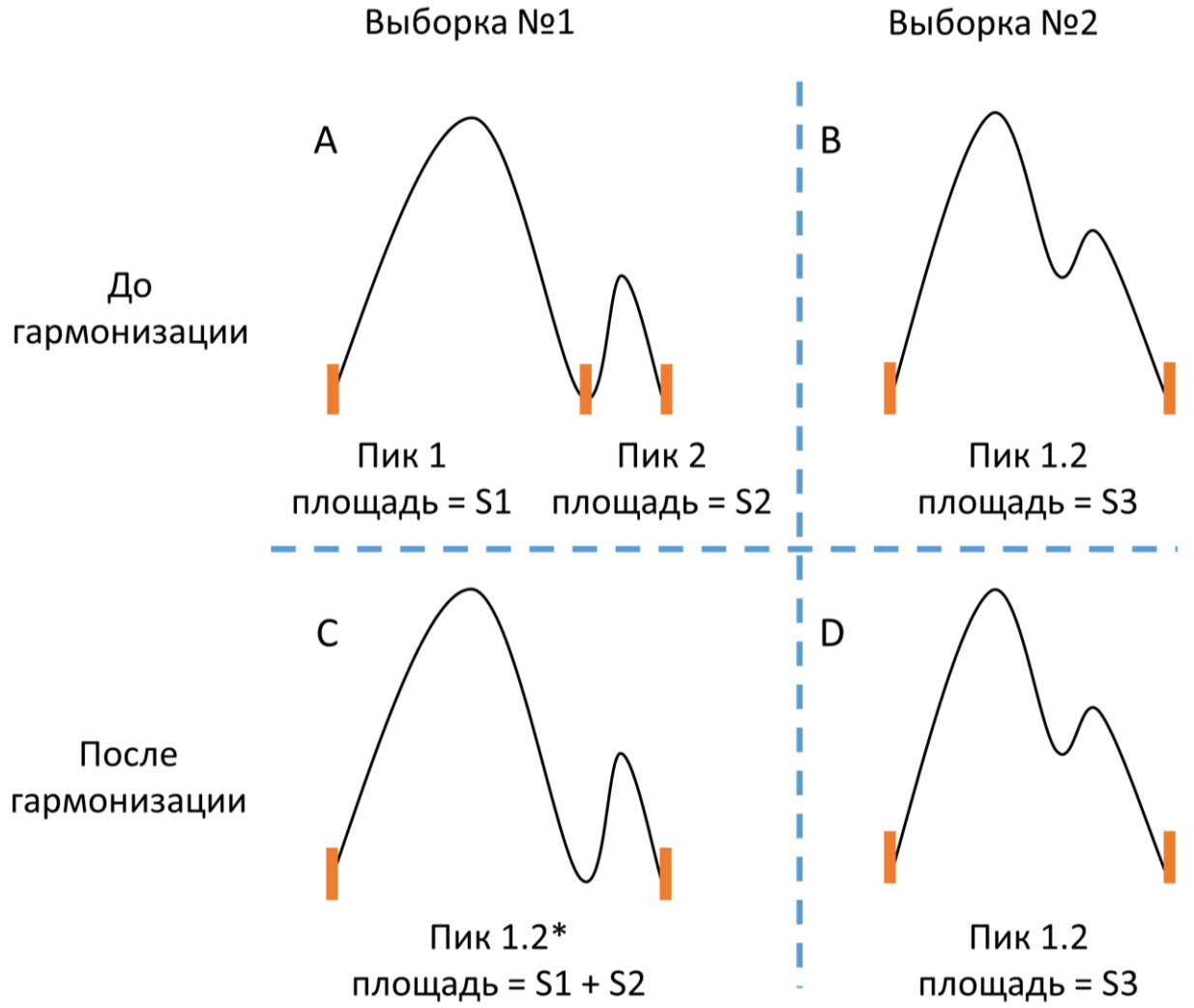


Рис. 14. Схема гармонизации гликомных пиков. A и B– участки исходных хроматограмм (до гармонизации) образцов из двух выборок. A – в выборке №1 пики 1 и 2 разделены. B - в выборке №2 пики 1 и 2 не разделены и образуют пик 1.2. C и D – участки хроматограмм после проведения гармонизации. C – площадь пика 1.2* равна сумме площадей двух исходных пиков 1 и 2. D – пик 1.2 соответствует гармонизированному пику 1.2*.

проиллюстрирована на Рис. 14. Для образцов выборки №1 измерены два пика

- «1» и «2», у образцов выборки №2 первые два пика оказались неразделенными, и они были объединены в один пик «1.2».

Гармонизацию гликомных профилей можно проводить методом изменения границ интегрирования пиков. В этом случае, хроматограммы каждого из образцов каждой из выборок требуют повторной ручной обработки, в ходе которой будут установлены единые для всех выборок границы интегрирования пиков. Однако, ручная обработка хроматограмм является трудоемким процессом. Можно оценить, что анализ 4,802 образцов в ручном режиме занимает 3 месяца работы квалифицированного сотрудника [37], использующего специальное ПО для обработки хроматограмм. Таким образом, встает задача разработки более эффективного метода гармонизации гликомных профилей.

Так как количественным признаком, отражающим уровень N-гликана, является площадь под соответствующим пиком, а в результате объединения двух пиков в один происходит объединение их площадей, то простым способом гармонизации пиков между выборками является суммирование площади соответствующих пиков. Мы предположили, что для гармонизации гликомных профилей между выборками 1 и 2 (Рис. 14) можно провести простое сложение площадей пиков «1» и «2» для всех образцов выборки №1 без необходимости дополнительного ручного анализа исходных данных хроматографии.

3.2.2. Реализация и валидация метода

Для реализации разработанного подхода была создана таблица соответствия гликомных пиков, полученных для исследуемых выборок (см Табл. 5), используя данные о биохимической структуре гликанов, входящих в состав гликомных пиков.

Табл. 5. Соответствие пиков на хроматограммах, полученных для исследуемых выборок. TwinsUK, EPIC Potsdam, PainOmics, SOCCS, SABRE - название выборок. Для каждой из выборок приведен набор пиков, измеренных по результатам хроматографии. Название признака - краткое наименование признака, полученного в результате гармонизации гликомных профилей.

TwinsUK	EPIC Potsdam	PainOmics	SOCCS	SABRE	Название признака
GP1	GP1	GP1	GP1	GP1	PGP1
GP2	GP2	GP2	GP2	GP2	PGP2
GP3	GP3	GP3	GP3	GP3	PGP3
GP4	GP4	GP4	GP4	GP4	PGP4
GP5	GP5	GP5	GP5	GP5	PGP5
GP6	GP6	GP6	GP6	GP6	PGP6
GP7	GP7	GP7	GP7	GP7	PGP7
GP8	GP8	GP8	GP8	GP8	PGP8
GP9	GP9	GP9	GP9	GP9	PGP9
GP10	GP10	GP10	GP10+GP11	GP10	PGP10
GP11	GP11	GP11	GP12	GP11	PGP11
GP12	GP12	GP12	GP13	GP12	PGP12
GP13	GP13	GP13	GP14	GP13	PGP13
GP14+GP15	GP14+GP15	GP14+GP15	GP15+GP16	GP14.15	PGP14
GP16	GP16	GP16	GP17+GP18	GP16	PGP15
GP17	GP17	GP17	GP19	GP17	PGP16
GP18	GP18	GP18	GP20+GP21	GP18	PGP17
GP19	GP19	GP19	GP22	GP19	PGP18
GP20+GP21	GP20+GP21	GP20+GP21	GP23+GP24	GP20.21	PGP19
GP22	GP22	GP22	GP25	GP22	PGP20
GP23	GP23	GP23	GP26	GP23	PGP21
GP24+GP25	GP24+GP25	GP24+GP25	GP27+GP28	GP24	PGP22
GP26	GP26	GP26	GP29	GP25	PGP23
GP27	GP27	GP27	GP30	GP26	PGP24
GP28	GP28	GP28	GP31	GP27	PGP25
GP29	GP29	GP29	GP32	GP28	PGP26
GP30	GP30	GP30	GP33	GP29	PGP27
GP31	GP31	GP31	GP34	GP30	PGP28
GP32	GP32	GP32	GP35	GP31	PGP29
GP33	GP33	GP33	GP36	GP32	PGP30
GP34	GP34	GP34	GP37	GP33	PGP31
GP35	GP35	GP35	GP38	GP34	PGP32
GP36	GP36	GP36	GP39	GP35	PGP33
GP37	GP37	GP37	GP40	GP36	PGP34
GP38	GP38	GP38	GP41	GP37	PGP35
GP39	GP39	GP39	GP42	GP38	PGP36

На основе Табл. 5 был составлен список из 36 гармонизированных гликомных пиков (см Доп. табл. 2) и схема гармонизации для каждой из выборок. Разработанный протокол гармонизации был апробирован на данных 35 случайно выбранных образцов из двух выборок: TwinsUK и PainOmics. Для этих образцов (согласно Табл. 5 и Доп. табл. 2) была проведена гармонизация пиков GP24 и GP25 в пик PGP22 двумя методами – ручным объединением пиков на хроматограмме (проведенной экспертом лаборатории Genos) и методом суммирования площадей. Результаты сравнительного анализа представлены на Рис. 15 и в Доп. табл. 1. Коэффициент корреляция Пирсона между значениями пика PGP22, полученного двумя методами, был более 0.9999, при этом разница между значением PGP22, полученным методом суммирования площадей и методом ручного объединения, наблюдалась лишь в восьмом значащем знаке.

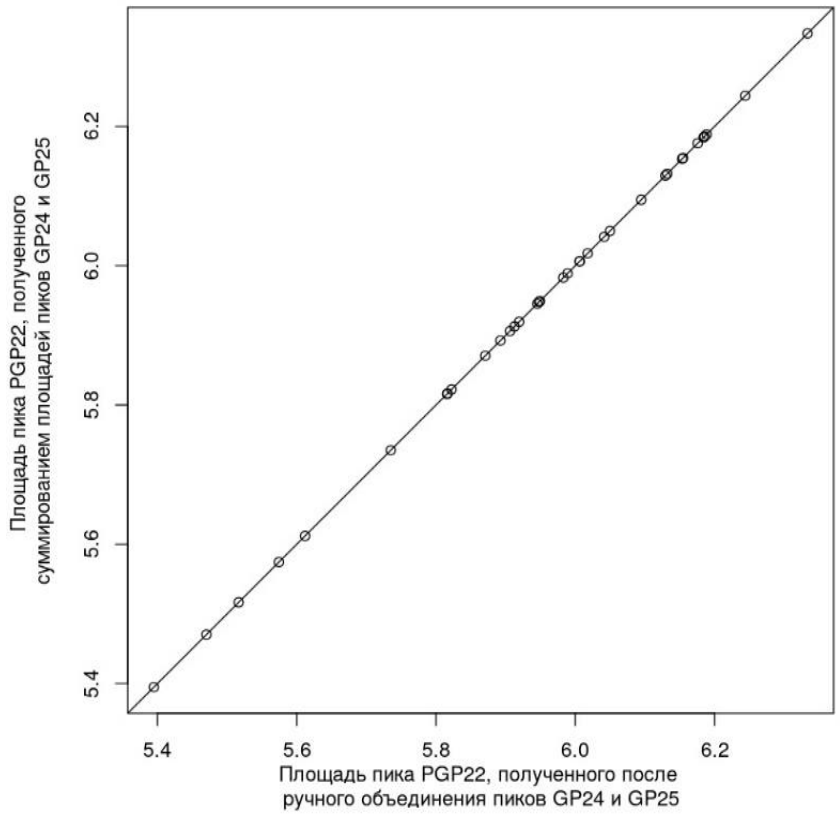


Рис. 15. Сравнение результатов гармонизации пиков GP24 и GP25 в пик PGP22. Гармонизация проведена для 35 образцов из выборок TwinsUK, и PainOmics. Ось X – площадь пика PGP22, гармонизированного методом ручного объединения пиков GP24 и GP25, ось Y – площадь пика PGP22, гармонизированного методом суммирования площадей. Значения площадей приведены на log10 шкале. Коэффициент корреляции Пирсона > 0.9999.

3.2.3. Краткое заключение

В данном разделе предложен и валидирован метод суммирования площадей. Данный метод позволяет получить единообразный набор гликомных признаков, измеренных методом СВЭЖХ в нескольких выборках. В виду простоты, низких временных затрат, и высокой точности в качестве метода гармонизации был выбран метод суммирования площадей. Разработанный метод и полученные результаты опубликованы в работе [124]

3.3. Подтверждение результатов ПГИА на независимых выборках

В первых двух полногеномных исследованиях ассоциаций N-гликома [33, 34] были найдены шесть локусов с полногеномно значимой ассоциацией. В данном исследовании была показана ассоциация 10 новых локусов, тем самым общее число локусов было увеличено с 6 до 16 . Однако ассоциация этих локусов не была подтверждена на независимых выборках, как того требуют стандарты, принятые в количественной генетике [135].

В данном разделе диссертации описаны результаты подтверждения ассоциаций 16 локусов на материале выборок EPIC-Potsdam, PainOmics, SOCCS и SABRE. Эти выборки не анализировались в предыдущих исследованиях, и на момент проводимого исследования они представляли собой самую большую коллекцию образцов геномных и гликомных данных. Гликомные профили в данных выборках были аннотированы согласно новым данным о структурах N-гликанов, вносящих основной вклад в формирование пиков на СВЭЖХ хроматограмме [91]. В этой аннотации для восьми пиков была предложена новая структура мажорного N-гликана.

3.3.1. Контроль качества N-гликомных данных и расчет производных признаков

Уровни N-гликанов в образцах плазмы крови участников исследований EPIC-Potsdam, PainOmics, SOCCS и SABRE были измерены методом СВЭЖХ.

Гликомные данные 10,004 образцов были гармонизированы с помощью предложенного в данной работе метода суммирования площадей. Гармонизированные данные 9,412 образцов прошли контроль качества (см. Табл. 6). Из них для 4,802 образцов были доступны геномные данные, прошедшие контроль качества: EPIC-Potsdam (N=2,192), PainOmics (N=1,874), SOCCS (N=459) и SABRE (N=277).

Были обновлены наборы производных признаков и формулы их расчета согласно аннотации 2020 года [91]. В итоге были рассчитаны 81 производный признак. Общее число признаков для образцов исследуемых выборок составило 117. Описание 117 признаков и формулы их расчета приведены в Доп. Табл. 3. Таким образом, в результате проведения контроля качества N-гликомных данных были подготовлены фенотипы образцов четырех независимых выборок для подтверждения ассоциации 16 локусов и для определения признаков, ассоциированных с ними.

Табл. 6. Результаты контроля качества данных гликома плазмы крови человека. N – число образцов с измеренным профилем N-гликозилирования; N КК - число образцов после контроля качества.

Выборка	N	N КК	Исключено образцов	% исключенных образцов	Число образцов с геномными данными
EPIC Potsdam	3,595	3,391	204	5.67%	2,192
PainOmics	3,242	3,018	224	6.91%	1,874
SOCCS	1,919	1,842	77	4.01%	459
SABRE	1,248	1,161	87	6.97%	277
Итого	10,004	9,412	592	5,91%	4,802

3.3.2. Подтверждение генетических ассоциаций 16 локусов на материале независимых выборок

Для каждого из 16 локусов был выбран маркирующий их ОНП. Для локусов *FUT3/FUT5/FUT6*, *FUT8*, *HNF1A*, *B3GAT1*, *MGAT5* и *SLC9A9*, чья ассоциация были показана в предыдущих ПГИА, были выбраны ОНП,

показавшие наименьшее P-value в исследовании [34], см. Табл. 1. Для 10 новых локусов были выбраны ОНП, показавшие наименьшее P-value ассоциации на выборке TwinsUK (см. Табл. 4 раздела 3.1). ОНП rs59111563, выбранный для локуса *ST6GAL1*, отсутствовал в четырех выборках. Поэтому для локуса *ST6GAL1* был выбран ОНП rs17775791, находящийся в неравновесии по сцеплению (r^2=0.99 в подгруппе образцов европейского происхождения выборки 1000 Genomes Project phase 3 version 5, аллель delT rs59111563 сцеплен с аллелем T rs17775791).

Анализ генетических ассоциаций 16 локусов проводился с использованием того же протокола, что и для поискового ПГИА, выполненного на выборке TwinsUK. Объединение результатов, полученных на четырех выборках, проводилось методом мета-анализа с фиксированными эффектами. Мета-анализ проводился с пользованием ПО METAL [129].

Напомним, что в данном анализе использовалась обновленная (по сравнению с данными выборки TwinsUK) панель гликомных признаков. Для получения обновленного списка признаков, ассоциированных с 16 локусами, тестировалась ассоциация локусов со всеми 117 N-гликомными признаками из обновленной панели. Ассоциация локуса считалась подтвержденной, если P-value ассоциации ОНП с хотя бы одним из 117 признаков было меньше использованного порога. Использовался уровень значимости P-value $< 0.05/(16 * 117) = 2.67 * 10^{-5}$, где 16 – число локусов, а 117 – число N-гликомных признаков. Объем выборки обеспечивал 95% статистическую мощность подтверждения ассоциаций для 15 локусов. Для локуса, расположенного на 22 хромосоме, 29.5 м.п.н. (rs140053014, интронный вариант гена *KREMEN1*) статистическая мощность репликации равнялась 9%. Это связано с тем, что данный ОНП прошел контроль качества только в выборке SOCCS (N = 472). В свою очередь, это может быть связано с низкой частотой минорного аллеля - 2% - ОНП rs140053014, представляющего данный локус.

В результате проведенного мета-анализа была подтверждена ассоциация 15 из 16 локусов (см. Табл. 7). Ассоциация локуса *KREMEN1* не была подтверждена. Это может объясняться недостаточной статистической мощностью (9%) подтверждения сигнала ассоциации. В этом случае, повторное подтверждение ассоциации локуса *KREMEN1* на выборке размером более 2,480 образцов позволит подтвердить ассоциацию данного локуса с 95% мощностью. В качестве альтернативного объяснения можно предположить, что ассоциация локуса *KREMEN1* может является ложноположительным результатом полногеномного анализа ассоциаций, выполненного на выборке TwinsUK. В этом случае ожидается, что данная ассоциация не будет подтверждена в последующих исследованиях генетического контроля уровней N-гликанов белков плазмы крови. На данный момент на основании полученных результатов нельзя сделать однозначный вывод об отсутствии или наличии роли локуса *KREMEN1* в контроле N-гликозилирования.

Для шести локусов наиболее ассоциированный признак остался тем же, что и в исследовании [34] (*MGAT5*) или в ПГИА из раздела 3.1 (*ST6GAL1*, *B4GALT1*, *IKZF1*, *IGH*/*TMEM121* и *SMARCB1*/*DERL3*/*CHCHD10*). Для девяти локусов признак с самой достоверной ассоциацией изменился. При этом для четырех локусов (*B3GAT1*, *FUT8*, *FUT3*/*FUT5*/*FUT6*, *HNF1A*) изменение признака имеет простое объяснение. Данные локусы в исследовании [34] показали наименьшее P-value ассоциации с признаками A4F2G4, A2F1G2, A2 и A3F1G3, измеренными после энзиматического десиалилирования N-гликома плазмы крови. В данном исследовании такой обработки не проводилось и такие признаки не измерялись.

В общей сложности 214 пар локус-признак показали ассоциацию с P-value $< 2.67 * 10^{-5}$. Сеть взаимосвязей между локусами и гликомными признаками визуализирована в разделе «Генная сеть регуляции N-гликозилирования», см. Рис. 17.

3.3.3. Краткое заключение

Мы подтвердили ассоциацию 15 из 16 локусов с уровнями N-гликанов плазмы крови с использованием независимых выборок. Высокий процент подтвержденных ассоциаций локусов (94%) свидетельствует о надежности результатов ПГИА, протоколов и стандартов, принятых сообществом, изучающим генетическую регуляцию гликозилирования белков. Полученные результаты подтверждают надежность найденных ранее ассоциаций и целесообразность проведения функциональных исследований 15 локусов. Ассоциация локуса гена *KREMEN1* с уровнями N-гликанов белков плазмы крови остается под вопросом. Полученные результаты опубликованы в работе [136]. В результате проведенной работы, мы показали, что популяционная изменчивость уровней N-гликанов, связанных с белками плазмы крови человека, контролируется как минимум 15 локусами генома, 9 из которых определены впервые.

Табл. 7. Подтверждение ассоциаций 16 локусов. В верхней части таблицы приведены результаты для шести локусов, ассоциация которых с гликомом была показана ранее [34]; в нижней части таблицы приведены результаты для десяти локусов, ассоциация которых показана впервые в данном исследовании. ОНП – ОНП, показавшие наименьшее P-value в исследовании [34] (верхняя часть таблицы), в данном исследовании (нижняя часть таблицы); Позиция – позиция ОНП (хромосома:п.н.) в геномных координатах (сборки генома GRCh37); ген – предложенный кандидатный ген в локусе; Эфф/Реф – эффекторный/референтный аллель; EAF – частота эффекторного аллеля; признак – признак с минимальным P-value ассоциации для данного ОНП; BETA (SE) – оценка эффекта эффекторного аллеля на признак и его стандартная ошибка; P – P-value ассоциации ОНП с признаком.

ОНП	Позиция	Ген	Эфф / Реф	EAF	Признак	BETA (SE)	*P*	Признак	BETA (SE)	*P*	EAF
			Huffman et al. 2011						**Данное исследование**		
rs1257220	2:135015347	*MGAT5*	A/G	0.26	Tetra-antennary glycans	0.19 (0.03)	1.80×10^{-10}	G4total, A4total	0.22 (0.02)	6.11×10^{-20}	0.26
rs4839604	3:142960273	*SLC9A9*	C/T	0.77	Tetrasialylated glycans	-0.22 (0.03)	3.50×10^{-13}	FBS2/FS2	-0.20 (0.03)	3.87×10^{-11}	0.80
rs7928758	11:134265967	*B3GAT1*	T/G	0.88	A4F2G4	0.23 (0.04)	1.66×10^{-08}	A4G4S3	0.36 (0.03)	6.43×10^{-27}	0.85
rs735396	12:121438844	*HNF1A*	T/C	0.61	A2F1G2	0.18 (0.03)	7.81×10^{-12}	G3Fa/G3total	0.21 (0.02)	4.91×10^{-20}	0.65
rs11621121	14:65822493	*FUT8*	C/T	0.43	A2	0.27 (0.03)	1.69×10^{-23}	FG3/G3total	-0.31 (0.02)	8.94×10^{-45}	0.42
rs3760776	19:5839746	*FUT3;FUT6*	G/A	0.87	A3F1G3	0.44 (0.04)	3.18×10^{-29}	G3Fa/G3total	0.48 (0.05)	3.85×10^{-23}	0.91
			Sharapov et al. 2019						**Данное исследование**		
rs186127900	1:25318225	*RUNX3*	G/T	0.99	FBG1n/G1n	-1.26 (0.12)	4.04×10^{-24}	FA2G2S2	1.24 (0.19)	1.16×10^{-10}	0.99
rs59111563[g]	3:186722848	*ST6GAL1*	Del/Ins	0.74	FG1S1/(FG1 + FG1S1)	0.34 (0.03)	1.09×10^{-26}	FG1S1/(FG1 + FG1S1)	0.49 (0.02)	8.60×10^{-97}	0.74
rs3115663	6:31601843	*HLA*	T/C	0.80	M9	0.26 (0.04)	7.65×10^{-11}	M9	0.15 (0.03)	1.63×10^{-07}	0.82
rs6421315	7:50355207	*IKZF1*	G/C	0.59	A2[6]BG1n	0.19 (0.03)	7.57×10^{-11}	A2[6]BG1n	0.23 (0.02)	1.19×10^{-27}	0.63
rs13297246	9:33128617	*B4GALT1*	G/A	0.83	FA2G2n	-0.26 (0.04)	4.11×10^{-12}	FA2G2n	-0.31 (0.03)	1.28×10^{-24}	0.83
rs3967200	11:126232385	*ST3GAL4*	C/T	0.88	A2G2S[3,6+3]2	-0.49 (0.04)	1.51×10^{-27}	G4S3/G4S4	0.63 (0.03)	1.20×10^{-106}	0.86
rs35590487	14:105989599	*IGH TMEM121*	C/T	0.77	FA2[3]G1n	-0.24 (0.03)	7.98×10^{-12}	FA2[3]G1n	-0.20 (0.03)	1.38×10^{-09}	0.75
rs9624334	22:24166256	*SMARCB1 DERL3 CHCHD10*	G/C	0.85	FA2[6]BG1n	0.28 (0.04)	8.38×10^{-12}	FA2[6]BG1n	0.31 (0.03)	7.15×10^{-26}	0.83
rs140053014	22:29550678	*KREMEN1*	Ins/Del	0.98	G3S2/G3S3	-0.67 (0.11)	4.05×10^{-10}	FA2[3]G1n	-0.68 (0.23)	0.0027	0.98
rs909674	22:39859169	*MGAT3*	C/A	0.27	FBS2/FS2	0.22 (0.03)	7.72×10^{-11}	FBn	0.22 (0.02)	1.88×10^{-20}	0.30

3.4. Приоритизация генов-кандидатов в найденных локусах

Знание локуса является отправной точкой в исследовании биологических механизмов, лежащих в основе влияния генетической изменчивости на формирование исследуемого признака. Сами по себе результаты ПГИА не позволяют сделать вывод о том, какие из генетических замен в локусе приводят к появлению ассоциации и функционирование каких генов изменяется в результате данных замен. Поскольку в локусе могут находится тысячи сцепленных генетических вариантов и последовательности десятков генов, число возможных механистических гипотез, обуславливающих найденные ассоциации, может быть очень велико. Проверка каждой из гипотез в функциональных лабораторных экспериментах является неэффективной в виду продолжительности и стоимости данных исследований. Поэтому после проведения ПГИА принято проводить количественно-генетические и биоинформатические анализы для отбора наиболее вероятных генов-кандидатов и гипотез, объясняющих найденные ассоциации. В данном разделе диссертации описаны результаты биоинформатического и количественно-генетического исследования 15 локусов генома с подтвержденной ассоциацией (см. раздел «Подтверждение результатов ПГИА на независимых выборках»).

Был определен набор ОНП, наиболее вероятно содержащий каузальные замены. Потенциальные последствия данных замен и гены, на которые они могут влиять были исследованы с использованием методов VEP [52] (включающего методы PolyPhen [137] и SIFT [138]), FATHMM-XF [53] и FATHMM-INDEL [54]. Для выявления клеточных типов и тканей, а также молекулярных путей, вовлеченных в формировании N-гликозилирования белков плазмы крови человека, использовались программное обеспечение и набор методов DEPICT [56].

3.4.1. Функциональная аннотация ОНП

Был составлен доверительный набор из 700 ОНП, замены в которых наиболее вероятно обуславливают ассоциации 15 найденных локусов. Из 700 ОНП 610 были одно-нуклеотидными заменами, 90 – короткими вставками/делециями.

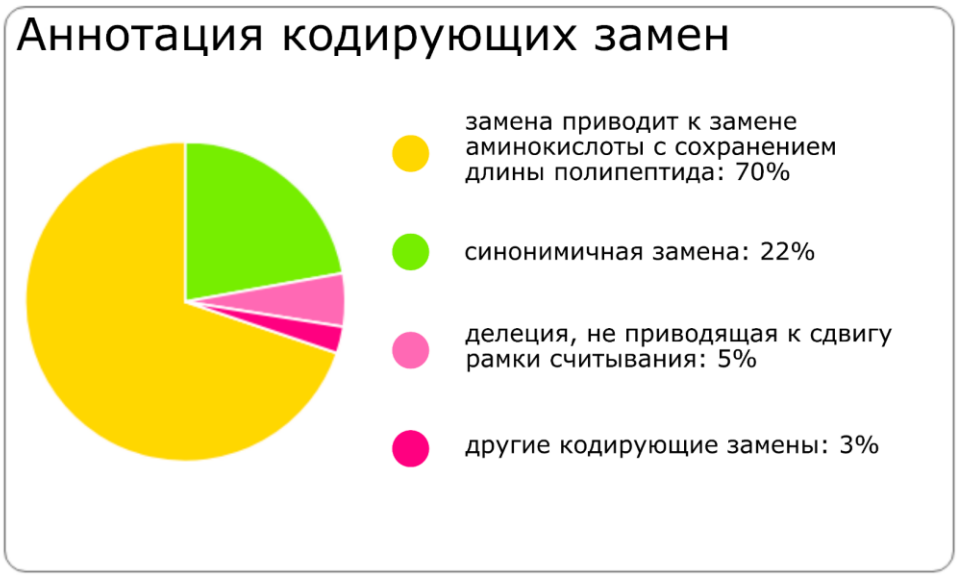


Рис. 16. Результаты функциональной аннотации ОНП, ассоциированных с гликомом плазмы крови. На круговой диаграмме слева показано распределение типов функциональных последствий замен, ассоциированных с гликомом плазмы крови. На круговой диаграмме справа показано распределение типов функциональных последствий замен, находящихся в кодирующей части.

Для определения возможных функциональных эффектов ОНП использовались программный пакет VEP и методы FATHMM-XF и FATHMM-

INDEL. Сводные результаты приведены на Рис. 16 и Табл. 8. Большая часть из 700 ОНП располагалась в интронных регионах генов (59%), либо в 3’ некодирующей области (16%). Примерно 3% замен находились в кодирующей части. Среди этих замен, 70% были несинонимичными заменами, 22% синонимичными, а 5% являлись делециями, не приводящими к сдвигу рамки считывания.

На основе полученных результатов приоритизированы 10 ОНП. Данные ОНП удовлетворяли хотя бы одному из двух условий: замена является несинонимичной и изменяет аминокислотную последовательность белка; замена классифицирована методами FATHMM-XF или FATHMM-INDEL как патогенная.

Восемь ОНП располагались в кодирующей части 21 транскрипта восьми генов –*PRRC2A, GPANK1, TMEM121, FUT8, NRTN, FUT6, DERL3* и *SYNGR1*. Одна замена располагалась в 3’-нетранслируемой области гена *BAG6* / после кодирующей области гена *PRRC2A* и еще одна замена находилась в 3’-нетранслируемой области гена *SYNGR1*. Таким образом на основе аннотации ОНП были приоритизированы девять генов.

Среди 10 приоритизированных ОНП стоит отметить замену rs17855739, расположенную на 19 хромосоме в позиции 5,831,840 п.н. Эта замена классифицирована методом FATHMM-XF как патогенная и является несинонимичной для пяти различных транскриптов гена *FUT6*. Замена Gag > Aag (частота ~ 12% согласно базе данных TOPMED [42]) в этой позиции приводит к замене p.Glu247Lys в аминокислотной последовательности фермента Fuc-TVI. Данный фермент является альфа 1,3-фукозилтрансферазой – ферментом, переносящим остаток фукозы к углеводному акцептору. При этом, замена p.Glu247Lys расположена в каталитическом домене фермента Fuc-TVI и она приводит к инактивации фермента. В гомозиготном состоянии она приводит к синдрому недостатка альфа 1,3-фукозилтрансферазы в плазме

крови [139], однако серьезных клинических последствий данная замена не имеет [140].

Табл. 8. Результаты функциональной аннотации замен в 15 локусах. Приведены замены. ОНП – идентификатор замены в базе данных dbSNP; Позиция – позиция замены в геноме (сборка генома GRCh38); Реф/Пат (Частота) – референсный и патогенный аллели (частота патогенного варианта согласно базе данных TopMED [42]; Описание замены – несинонимичная замена с указанием аминокислотной замены и ее расположения; ген – название гена, в кодирующей области которого лежит замена. XF – классификация замены методом FATHMM-XF; INDEL – классификация делеции/инсерции методом FATHMM-INDEL. Нейтр. – нейтральная замена, Пат. – патогенная замена.

ОНП	**Позиция (GRCh38)**	**Реф/Пат (Частота)**	**Описание замены**	**Ген**	**XF**	**INDEL**
rs115201868	6:31636814	C/T (16%)	Несинонимичная p.Pro2006Ser	*PRRC2A*	Нейтр.	-
rs5875328	6:31639125	dupAAGA/- (16%)	Расположена в 3’ нетранслируемой области гена *BAG6* После кодирующей области гена *PRRC2A*	*BAG6 PRRC2A*	-	Пат.
rs3130618	6:31664357	C/A (16%)	Несинонимичная p.Arg41Leu	*GPANK1*	Пат.	-
rs2229677	14:65561728	A/C (29%)	Несинонимичная p.Gln55His	*FUT8*	Нейтр.	-
rs10569304	14:105529713	CGCCGCCGCCGC/- (39%)	Делеция p.Pro296_Pro299del	*TMEM121*	-	Нейтр.
rs79744308	19:5827754	G/A (5%)	Несинонимичная p.Ala59Thr	*NRTN*	Пат.	-
rs17855739	19:5831829	G/A (12%)	Несинонимичная p.Glu247Lys	*FUT6*	Пат.	-
rs3177243	22:23837735	G/C (17%)	Несинонимичная p.Phe149Leu	*DERL3*	Пат.	-
rs149306472	22:39381818	ACA/- (25%)	Делеция p.Pro202_Thr203insSer	*SYNGR1*	-	Пат.
rs7423	22:39385424	C/A (36%)	Расположена в 3’ нетранслируемой области гена *SYNGR1*	*SYNGR1*	Пат.	-

3.4.2. Анализ биологических путей и тканеспецифичной экспрессии

Для идентификации биологических путей, влияющих на гликозилирование, анализа тканеспецифичной экспрессии и приоритизации генов на основе данных анализов, использовался метод DEPICT [56]. Для анализа были выбраны 15 локусов из Табл. 7. Было обнаружено, что в найденных локусах расположены гены, экспрессирующиеся преимущественно в плазмоцитах, в околоушных железах, слюнных железах, в клетках, продуцирующих антитела и B-лимфоцитах (обогащение на уровне

значимости ожидаемой доли ложных отклонений – FDR<0.05). На основе предсказанной функциональной роли генов и вовлеченности генов в те или иные биологические пути, метод DEPICT приоритизировал гены *FUT3*, *DERL3* и *FUT8* для трех локусов (на хромосоме 19 в 58 м.п.н, на хромосоме 22 в 24 м.п.н. и на хромосоме 14 в 65/66 м.п.н.

3.4.3. Анализ плейотропных эффектов локусов на уровни экспрессии близлежащих генов

Для определения генов, чей уровень экспрессии может опосредовать эффекты найденных локусов на гликом плазмы крови, был применен метод Менделевской рандомизации (SMR) с последующим анализом гетерогенности (HEIDI) [55]. SMR/HEIDI анализ выявляет гены, чей уровень экспрессии в определенной ткани контролируется теми же ОНП в локусе, что и исследуемый признак. Анализ был проведен для 12 из 15 локусов с подтвержденной ассоциацией. Данный анализ не был проведен для трех локусов. Локусы *MGAT5* и *SLC9A9* не показали полногеномно значимую ассоциацию в ПГИА на выборке TwinsUK (см. Табл. 4), а локус *HLA* был исключен из анализа согласно рекомендациям авторов метода SMR/HEIDI [55] в виду протяженного неравновесия по сцеплению в данном локусе. Для проведения SMR/HEIDI анализа были использованы данные GTEx [50], Westra [105] и CEDAR [51]. В общей сложности в анализ было включено 20,448 уровней экспрессии генов. Для дальнейшего анализа были отобраны только уровни экспрессии генов в тканях и типах клеток, наиболее близких к клеткам, секретирующим антитела или к печени: CD19+ B-лимфоциты, кровь и печень. Результаты теста SMR показали, что семь локусов, ассоциированных с уровнями гликанов, также ассоциированы с уровнями экспрессии 12 генов на уровне значимости $P_{SMR} \leq 0.05/20{,}448 = 2.4 * 10^{-6}$ (см Табл. 9).

Для уровней экспрессии 12 генов был проведен анализ HEIDI. В результате для уровней экспрессии четырех генов тест HEIDI не смог отвергнуть гипотезу о плейотропии ($P_{HEIDI} > 0.05$) - экспрессии генов

ST6GAL1, *CHCHD10* и *TMEM121* в периферической крови и экспрессии гена *MGAT3* в В-лимфоцитах. Таким образом, уровень экспрессии данных генов может опосредовать найденные ассоциации этих локусов с уровнями N-гликанов плазмы крови. Для уровня экспрессии гена *B3GAT1* в периферической крови не была принята или опровергнута гипотеза ($0.001 < P_{HEIDI} < 0.05$) о плейотропии, что говорит о возможном плейтропном генетическом контроле локусом экспрессии данного гена и уровней N-гликанов белков плазмы крови.

Табл. 9. Результаты анализа SMR/HEIDI. В таблице приведены результаты анализа eQTL эффектов локусов, ассоциированных с гликомом плазмы крови человека. В таблице приведены результаты для уровней экспрессии генов с $P_{SMR} \leq 2.4 * 10^{-6}$. Локус – ОНП из таблицы Табл. 4, представляющий локус; Геномные координаты – координаты ОНП, показавшего наименьшее P-value ассоциации в данном локусе; Гликан – гликанный признак, наиболее ассоциированный с локусом (см Табл. 4); Транскрипт – условное название транскрипта (изоформы гена); Ген – название гена; Ткань – ткань, в которой был измерен уровень экспрессии гена; данные – источник транскрипционных данных; beta SMR – оценка размера эффекта уровня экспрессии гена на N-гликомный признак; P_{SMR} – P-value SMR теста; P_{HEIDI} – P-value HEIDI теста.

Локус	**Геномные координаты**	**Гликан**	**Транскрипт**	**Ген**	**Ткань**	**Данные**	**beta SMR**	**P_{SMR}**	**P_{HEIDI}**
rs3967200	11:126232385	PGP17	ILMN_1762312	*FOXRED1*	Переф. кровь	Westra	-1,434	1,27E-15	2,30E-18
rs3967200	11:126232385	PGP17	ILMN_1730082	*RPUSD4*	Переф. кровь	Westra	-0,538	3,84E-13	2,04E-07
rs59111563	3:186722848	PGP41	ILMN_1756501	*ST6GAL1*	Переф. кровь	Westra	0,549	1,58E-11	1,69E-01
rs1169303	12:121436376	PGP30	ILMN_1737818	*C12ORF43*	Переф. кровь	Westra	-1,783	1,59E-09	2,52E-06
rs35590487	14:105989599	PGP62	ENSG00000184986.6	*TMEM121*	Переф. кровь	GTEx	2,446	2,31E-08	5,67E-01
rs909674	22:39859169	PGP56	6420500_39887987_39888036	*MGAT3*	В лимфоциты	CEDAR	1,414	2,64E-08	4,16E-01
rs1866767	11:134274763	PGP33	ILMN_1761093	*B3GAT1*	Переф. кровь	Westra	1,415	2,78E-08	5,24E-03
rs1169303	12:121436376	PGP30	ILMN_1674811	*OASL*	Переф. кровь	Westra	-0,997	8,89E-08	2,64E-05
rs35590487	14:105989599	PGP62	ILMN_1694432	*CRIP2*	Переф. кровь	Westra	0,753	1,86E-07	
rs35590487	14:105989599	PGP62	ILMN_1653553	*C14ORF80*	Переф. кровь	Westra	-0,663	9,26E-07	
rs9624334	22:24166256	PGP63	ILMN_1740170	*CHCHD10*	Переф. кровь	Westra	0,625	1,66E-06	1,93E-01

3.4.4. Определение возможных плейотропных эффектов найденных локусов на мультифакторные признаки и заболевания человека

Найденные локусы были проверены на наличие возможных плейотропных эффектов на признаки и заболевания человека. Информация об ассоциациях 700 ОНП была извлечена из базы данных “PhenoScanner” [104]. В общей сложности 459 ОНП из 15 локусов показали ассоциацию с 264 признаками и заболеваниями человека на уровне значимости P-value $< 5 * 10^{-8}$.

Четыре локуса (*IKZF1*, *FUT8*, *HNF1A*, *RUNX3*) показали ассоциацию с признаками, отражающими уровень красных и белых кровяных телец и тромбоцитов [141]. Локус *FUT6*/*FUT5*/*FUT3* показал ассоциацию с возрастной дегенерацией желтого пятна [142, 143]. Локусы *HNF1A* и *ST3GAL4* были ассоциированы с уровнями различных липидов в крови [144, 145]. Локусы *HNF1A* и *ST3GAL4* показали ассоциацию с уровнями С-реактивного белка в крови [146, 147]. Локус *HNF1A* также был ассоциирован с риском развития диабета 2 типа [148], с уровнем γ-глютамил трансферазы [149]. Локус *FUT8* на хромосоме 14 в 65/66 м.п.н. был ассоциирован с возрастом начала менструаций [150]. Локус *MGAT3* на хромосоме 22 39 м.п.н. был ассоциирован с ростом человека в взрослом возрасте [151]. Локус *IGH*/*TMEM121* показал ассоциацию с риском развития ревматоидного артрита [152].

Таким образом, найденные в данной работе локусы показали ассоциацию с признаками и заболеваниями человека, в том числе с аутоиммунными заболеваниями, маркерами воспаления и маркерами заболеваний печени. Отметим, что результаты данного анализа указывают лишь на пересечение сигналов генетических ассоциаций уровней N-гликанов белков плазмы крови и данных признаков. Строгая проверка гипотезы о

наличии плейтропного эффекта (напр. методами SMR/HEIDI) не проводилась, поскольку это выходит за рамки поставленной цели и задач.

3.4.5. Предложенные гены-кандидаты

В результате проведенных анализов для 15 локусов были приоритизированы 24 гена-кандидат (см. колонку «Ген-кандидат», Табл. 10). В данном разделе мы приводим описание этих генов и предложенные гипотезы об их роли в регуляции N-гликозилирования белков плазмы крови человека. В результате обобщения полученных результатов мы приоритизировали 18 генов (см. колонку «Приоритизированные гены», Табл. 10).

В восьми локусах (*MGAT5*, *MGAT3*, *FUT3*/*FUT6*, *FUT8*, *ST6GAL1*, *ST3GAL4*, *B4GALT1*, *B3GAT1*) из пятнадцати были приоритизированы гены, кодирующие ферменты - гликозилтрансферазы, участвующие в биосинтезе N-гликанов. За исключением локуса *B3GAT1*, набор N-гликомных признаков, ассоциированных с данными локусами, соответствовал ферментативной активности данных ферментов.

В локусе на 2-ой хромосоме, 125 м.п.н., был приоритизирован ген *MGAT5*, кодирующий фермент GnT-V - альфа-1,6-маннозилгликопротеин 6-бета-N-ацетилглюкозаминилтрансферазу. Данный фермент переносит остаток N-ацетилглюкозамина к маннозе в составе N-гликана, при этом образуется три или тетраантеннарный N-гликан. Локус с геном *MGAT5* показал ассоциацию с гликомными признаками, отражающими уровень три- и тетра- антеннарных гликанов.

Табл. 10. Сводная таблица результатов биоинформатического функционального анализа 15 реплицированных локусов. Для каждого локуса был предложен ген(ы)-кандидат(ы) на основе ряда критериев: Ближ. ген – ген, чья последовательность наиболее близко расположена к ОНП, наиболее ассоциированному в данном локусе; Ген-кандидат – предложенные гены-кандидаты; Кандидатные ОНП – приоритизированные ОНП; eQTL – уровень экспрессии данного гена опосредует найденную ассоциацию локуса с гликомом; DEPICT – ген был приоритизирован методов DEPICT; Экспер. данные - для данного гена были проведены функциональные исследования, подтвердившие его роль в регуляции гликозилирования; Биосинтез N-гликанов – фермент, участвующий в биосинтезе гликанов, закодированный данным геном; Приоритиз. гены – гены, приоритизированные нами из списка генов-кандидатов.

Локус	Ближ. ген	Ген-кандидат	Кандидатные ОНП	eQTL	DEPICT	Экспер. данные	Биосинтез N-гликанов	Приоритиз. гены
Ранее известные локусы								
2:135015347	*MGAT5*	*MGAT5*					Альфа-1,6-маннозилгликопротеин 6-бета-N-ацетилглюкозаминилтрансфераза	*MGAT5*
3:142960273	*SLC9A9*	*SLC9A9*						*SLC9A9*
11:134274763	*B3GAT1*	*B3GAT1*		Кровь			Галактозилгалактозилксилозилпротеин-3-бета-глюкуронозилтрансфераза типа 1	*B3GAT1*
12:121436376	*HNF1A*	*HNF1A*				[33]		*HNF1A*
14:66011184	*FUT8*	*FUT8*	rs2229677		+		Альфа-(1, 6)-фукозилтрансфераза	*FUT8*
19:5828064	*NRTN*	*FUT3*			+		Галактозид-3(4)-L-фукозилтрансфераза	*FUT6*; *FUT3*
		FUT6	rs17855739				Альфа-(1,3)-фукозилтрансфераза	
		FUT5					Альфа-(1,3)-фукозилтрансфераза	
		NRTN	rs79744308					
Новые локусы, найденные в данной работе								
1:25318225	*RUNX3*	*RUNX3*						*RUNX3*
3:186722848	*ST6GAL1*	*ST6GAL1*		Кровь			Бета-галактозид альфа-2,6-сиалилтрансфераза 1	*ST6GAL1*
6:31601843	*PRRC2A*	*PRRC2A*	rs115201868 rs5875328					
		GPANK1	rs3130618					

		BAG6	rs5875328					
7:50355207	*IKZF1*	*IKZF1*				[153]		*IKZF1*
9:33128617	*B4GALT1*	*B4GALT1*					Бета-1,4-галактозилтрансфераза 1	*B4GALT1*
11:126232385	*ST3GAL4*	*ST3GAL4*					ЦМФ-N- ацетилнейрамин-бета-галактозамид-альфа-2,3-сиалилтрансфераза	*ST3GAL4*
14:105989599	*C14orf80*	*TMEM121*	rs10569304	Кровь				*TMEM121; IGH*
		IGH	-					
22:24166256	*SMARCB1*	*DERL3*	rs3177243		+			*SMARCB1; DERL3; CHCHD10*
		SMARCB1						
		CHCHD10		Кровь				
22:39859169	*MGAT3*	*MGAT3*		CD19+ клетки (B лимфоциты)			Бета-1,4-маннозил-гликопротеин 4-бета-N-ацетилглюкозаминилтрансфераза	*MGAT3*
		SYNGR1	rs149306472 rs7423					

В локусе на 22-ой хромосоме, 39 м.п.н., приоритизирован ген *MGAT3*, кодирующий N-ацетилглюкозаминилтрансферазу GnT-III - бета-1,4-маннозил-гликопротеин 4-бета-N-ацетилглюкозаминилтрансферазу. Данный фермент переносит остаток N-ацетилглюкозамина к маннозе в составе N-гликана в определенную позицию, при этом образуется рассечение остова. Было показано наличие плейтропного эффекта данного локуса как на уровни N-гликанов, так и на уровень экспрессии гена *MGAT3* в CD19+ клетках (B-лимфоцитах). По совокупности результатов, ген *MGAT3* представляется наиболее вероятным геном-кандидатом в этом локусе.

В локусе на 14-ой хромосоме, 66 м.п.н., приоритизирован ген *FUT8*. Этот ген кодирует фермент Fuc-TVIII - альфа-(1, 6)-фукозилтрансферазу. Данный фермент переносит остаток фукозы к N-ацетилглюкозамину, находящемуся в остове N-гликана и, тем самым, отвечает за фукозилирование остова N-гликанов. Локус показал ассоциацию с признаками, отражающими фукозилирование остова N-гликанов. Стоит отметить, что локусы *FUT8* и *MGAT3* показали ассоциацию с признаками FBS2/(FS2+FBS2) и FBS2/FS2, отражающими представленность рассечения остова у биантеннарных гликанов с фукозилированием остова, что отражает известный факт интерференции активностей ферментов Fuc-TVIII и GnT-III [154].

В локусе на 19-ой хромосоме, 5.8 м.п.н., приоритизированы гены *FUT6, FUT5, FUT3, NRTN*. Ген *NRTN* кодирует нейротропический фактор, регулирующие выживание и функционирование нейронов. Гены *FUT6* и *FUT3/FUT5* кодируют ферменты Fuc-TVI и Fuc-TIII - фукозилтрансферазы 6 и 3 соответственно, переносящие остаток фукозы с ГДФ-фукозы к N-ацетилглюкозамину путем образования альфа-1,3(4)-гликозидной связи. Эти ферменты ответственны за антеннарное фукозилирование N-гликанов. В данной работе было показано, что данный локус был ассоциирован с антеннарным фукозилированием три- и тетраантеннарных гликанов. Таким образом гены *FUT3/FUT5/FUT6* были выбраны в качестве кандидатных для данного локуса. Более того, в гене *FUT6*

(расположенном на 19 хромосоме 58 м.п.н.) расположен ОНП rs17855739. Этот ОНП кодирует замену G > A (частота аллеля A ~ 12% в популяциях человека согласно базе данных TopMED), которая приводит к замещению отрицательно заряженной глутаминовой кислоты на положительно заряженный лизин в 247 позиции (замена p.Glu247Lys). Эта замена расположена в каталитическом домене фермента Fuc-TVI и она приводит к инактивации фермента. Таким образом, данный вариант может иметь функциональное влияние на гликозилирование белков плазмы крови человека. Стоит отметить, что предполагается, что гены *FUT3, FUT5* и *FUT6* произошли от общего предкового гена в результате двух дупликации [155]. При этом экспрессия гена *FUT5* на уровне транскрипции и трансляции в организме человека гораздо слабее по сравнению с *FUT3* и *FUT6* [156]. В итоге, в данном локусе мы приоритизируем гены *FUT6* и *FUT3* как наиболее вероятные кандидаты.

В локусе на 3-ей хромосоме, 186 м.п.н., приоритизирован ген *ST6GAL1*. Результаты тестов SMR/HEIDI показали наличие возможного плейтропного эффекта данного локуса на уровень экспрессии гена *ST6GAL1* в клетках крови. Ген *ST6GAL1* кодирует альфа-2,6-сиалилтрансферазу 1. Данный фермент катализирует образование альфа-2,6-гликозидной связи между сиаловой кислотой и N-ацетилглюкозамином, связанным с галактозой, в составе N-гликана. Локус *ST6GAL1* показал ассоциацию с уровнями моно- и дисиалилированных N-гликанов и их предшественников.

В локусе на 11-ой хромосоме, 126 м.п.н., приоритизирован ген *ST3GAL4*. Ген *ST3GAL4* кодирует фермент альфа-2,3-сиалилтрансферазу, переносящую остаток сиаловой кислота. Данный локус показал ассоциацию с уровнями различных сиалилированных N-гликанов.

В локусе на 9-ой хромосоме, 33 м.п.н., приоритизирован ген *B4GALT1*. Ген *B4GALT1* кодирует фермент галактозилтрансферазу, присоединяющий галактозу к различным субстратам, в том числе и к N-ацетилглюкозамину. Локус *B4GALT1* был ассоциирован с уровнями галактозилированных биантеннарных N-гликанов и их

предшественников. Также известно, что ряд мутаций в гене *B4GATL1* приводит к врожденному заболеванию гликозилирования [157].

В локусе на 11-ой хромосоме, 134 м.п.н., был приоритизирован ген *B3GAT1*, кодирующий фермент галактозилгалактозилксилозилпротеин-3-бета-глюкуронозилтрансфераза 1 типа. Этот фермент катализирует перенос глюкуроновой кислоты в ходе биосинтеза HNK-1 эпитопа. Этот эпитоп экспрессируется на лимфоцитах, однако его присутствие на белках плазмы крови до определенного момента не было известно. Ассоциация данного локуса с уровнями N-гликанов белков плазмы крови была показана в работе [34]. В этой же работе исследователи обнаружили присутствие глюкуроновой кислоты в N-гликоме плазмы крови, что может объяснить ассоциацию данного локуса.

В семи других локусах приоритизированные гены-кандидаты не являлись генами гликозилтрансфераз. В локусе на 22-ой хромосоме 39 м.п.н. были приоритизированы три гена – *SMARCB1*, *DERL3*, *CHCHD10*. В кодирующей последовательности гена *SMARCB1* расположен самый сильный сигнал ассоциации в данном локусе. Ген *SMARCB1* кодирует белок, являющийся частью комплекса hSWI/SNF – ремоделера хроматина. Продукт гена *SMARCB1* играет важную роль в ингибировании канцерогенеза, пролиферации и дифференциации клеток [158]. Однако метод DEPICT показал, что в данном локусе возможным каузативным геном является *DERL3*, кодирующий фермент, участвующий в деградации люминальных гликопротеинов с некорректной третичной структурой, происходящей в эндоплазматическом ретикулуме [159]. Более того, анализ последствий замен из доверительного набора методом FATHMM-XF показал, что в данном локусе расположен патогенный вариант rs3177243, расположенный в кодирующей последовательности гена *DERL3*. В тоже время тесты SMR/HEIDI показали, что ассоциация данного локуса с гликомом плазмы крови может быть опосредована плейотропным эффектом локуса на уровень экспрессии гена *CHCHD10*. Ген *CHCHD10* кодирует митохондриальный белок, встречающийся в фибриллах крист митохондрий. Ген *CHCHD10* сильнее всего экспрессируется в

печени и сердечной мышце и менее всего в селезенке [106]. Ранее, непосредственное участие митохондриальных белков в процессах гликозилирования не было известно, однако в 2017 году была опубликована работа, показавшая роль фрагментации митохондрий и числа контактов эндоплазматического ретикулума с митохондриями в представленности сиалилированных гликанов на поверхности глиобластомных клеток, что в свою очередь влияло на узнавание лимфоцитами клеток глиобластомы [160]. Таким образом, в локусе были выбраны три возможных гена кандидата – *SMARCB1*, *DERL3* и *CHCHD10*. Стоит отметить, что этот локус и локус *MGAT3* были ассоциированы со сходными по структуре гликанами (гликанов с рассечением и фукозилированием остова), что может говорить о том, что данный процесс может регулироваться совместно генами *MGAT3* и *SMARCB1*/*DERL3*/*CHCHD10*.

Локус на хромосоме 14 в 105 м.п.н. содержит кластер генов *IGH*, кодирующие тяжелые цепи иммуноглобулинов. Иммуноглобулин G (IgG) является наиболее представленным N-гликопротеином плазмы крови человека [3]. Однако анализ SMR/HEIDI показал, что ассоциация данного локуса с гликомом плазмы может опосредоваться регуляцией экспрессии гена *TMEM121*. Данный ген кодирует трансмембранный белок 121, и экспрессируется в тканях сердечной мышцы, поджелудочной железы, печени и скелетных мышцах. Более того, анализ эффекта замен VEP показал, что вариант rs35590487, показавший самую сильную ассоциацию гликомом плазмы в данном локусе, находится в сильном неравновесии по сцеплению с вариантом rs10569304 (r^2=0.95, данные исследования 1000 геномов, 503 образца европейского происхождения) – делецией, расположенной в экзоне гена *TMEM121*. Аннотация алгоритмами PolyPhen и SIFT показала, что замена в ОНП rs10569304 может оказывать эффект на транскрипт гена *TMEM121*. Таким образом, гены *IGH* и *TMEM121* были выбраны в качестве генов-кандидатов.

В локусе на 3-ей хромосоме, 142 м.п.н. был приоритизирован ген *SLC9A9*. Ген *SLC9A9* кодирует Na+/H насос, который предположительно регулирует уровень pH в Аппарате Гольджи (АГ). Гликозилирование белков в происходит в

АГ, и, согласно имеющимся данным, этот процесс чувствителен к изменению pH [161]. Процессы, происходящие в АГ, влияют на формирование гетеродимерных комплексов, осуществляющих гликозилирование [162]. В работе [107] было показано, что увеличение pH в АГ может нарушить терминальное N-гликозилирование (включая сиалилирование) из-за неверной локализации гликозилтрансфераз. В согласии с этой гипотезой, в работе [34] локус *SLC9A9* показал ассоциацию с уровнем тетра-сиалилированных N-гликанов, а в данной работе – с уровнем сиалилированных N-гликанов.

В локусе на 12-ой хромосоме, 121 м.п.н., расположено три гена - *HNF1A, C12orf43* и *OASL*. Подробное функциональное исследование данного локуса, проведенное в работе [33], показало, что ген *HNF1A*, кодирующий фактор транскрипции гепатоцитов, регулирует экспрессию большинства генов, кодирующих фукозилтрансферазы - *FUT3, FUT5, FUT6, FUT8, FUT10, FUT11* - в клеточной линии HepG2, полученной из клеток печени. В том же исследовании было показано, что *HNF1A* регулирует экспрессию генов, кодирующих ключевые ферменты для синтеза ГДФ-фукозы – субстрата для фукозилтрансфераз, что в совокупности говорит о ключевой роли гена *HNF1A* в процессах фукозилирования гликанов. Таким образом, ген *HNF1A* был выбран в качестве кандидатного гена в данном регионе.

Для локуса на 7 хромосоме в 50 м.п.н. был приоритизирован ген *IKZF1*. Ранее в работе [64] была показана ассоциация данного локуса с гликозилированием IgG, в которой авторы предположили, что ген *IKZF1* является кандидатным для данного локуса. Ген *IKZF1* кодирует ДНК-связывающий белок Ikaros, - регулятор транскрипции, связанный с ремоделированием хроматина. Интересно, что локус *IKZF1* показал ассоциацию с уровнями N-гликанов белков плазмы крови с фукозилированием остова, с которыми был ассоциирован локус *FUT8* (см. Рис. 17). *IKZF1* рассматривается как важный регулятор дифференциации лимфоцитов [163, 164]. Поскольку клетки, секретирующие IgG, являются производными лимфоцитов, ген *IKZF1* был выбран в качестве кандидатного в данном локусе и

была выдвинута гипотеза о его роли в регуляции фукозилирования остова N-гликанов IgG путем регуляции экспрессии гена *FUT8*. Более того, в исследовании [153], выполненном в соавторстве с соискателем, было показано, что нокдаун гена *IKZF1* в IgG секретирующих клетках MATAT6 приводит к более чем трехкратному увеличению экспрессии *FUT8* и увеличению уровня фукозилирования секретируемого IgG. Таким образом, выдвинутая нами гипотеза нашла свое подтверждение.

В локусе на 1-ой хромосоме, 25 м.п.н., был приоритизирован ген *RUNX3*. Данный ген кодирует Runt-домен-содержащий белок - фактор транскрипции, который также, как и ген *IKZF1* [164] играет важную роль в созревании и дифференцировке В-лимфоцитов.

Приоритизация генов в локусе *HLA* (локус главного комплекса гистосовместимости человека) на 6-ой хромосоме 25-32 м.п.н. была проведена, однако ее результаты имеют высокий шанс быть ложноположительными. С точки зрения количественной генетики мультифакторных признаков человека локус *HLA* является уникальным [165]. Данный локус имеет самую большую плотность генов в геноме человека; он имеет высочайшую степень полиморфности на нуклеотидном уровне; аллели данного локуса находятся в высоком неравновесии по сцеплению на протяжении всего локуса длинной в 8 м.п.н. Все это не позволяет проводить тонкое картирование и приоритизацию генов в данном локусе общепринятыми методами. Однако кратко результаты приоритизации генов все же будут описаны. В этом локусе были приоритизированы три гена: *PRRC2A*, *GPANK1* и *BAG6*. Все три гена расположены в регионе *HLA*. Информация о данных генах достаточно скудна и ниже приведена информация из базы данных GeneCards. В гене *PRRC2A* присутствуют микросателлитные повторы, которые связаны с возрастом начала диабета I типа, и, возможно, вовлечены в воспалительный процесс разрушения бета-клеток поджелудочной железы во время развития диабета. Ген *BAG6* кодирует ядерный белок, вовлеченный в процесс апоптоза.

Таким образом, в результате проведённой работы мы предложили новые гены-кандидаты, вовлеченные в процесс N-гликозилирования - гены регуляторов транскрипции (*IKZF1*, *SMARCB1* и *RUNX3*), деградации гликопротеинов (*DERL3*), тяжелой цепи иммуноглобулинов (*IGH*) и гены с неизвестной функцией (*TMEM121* и *CHCHD10*).

3.5. Генная сеть регуляции N-гликозилирования

В данной работе был проведен анализ генетических ассоциаций с использованием самой крупной коллекции образцов с измеренными геномами и N-гликомами плазмы крови. Была показана достоверная ассоциация 15 локусов со 116 из 117 признаков. Суммарно 214 пар локус-признак показали достоверную ассоциацию (см. раздел Подтверждение результатов ПГИА на независимых выборках). Эти данные были использованы для реконструкции генной сети регуляции уровней N-гликанов белков плазмы крови (см. Рис. 17). Данная сеть визуализирует ассоциацию между найденными локусами и уровнями N-гликанов белков плазмы крови.

Для построения сети была проведена классификация гликомных признаков. Признаки были классифицированы на четыре группы согласно ткани, секретирующей N-гликопротеины в плазму крови. Первая группа – признаки, отражающие уровни N-гликанов иммуноглобулинов (A, G, D, E, M), секретируемых клетками лимфоцитарного ряда - B-лимфоцитами, плазмобластами и плазмоцитами. Вторая группа - признаки, отражающие уровни N-гликанов белков (трансферрин, гаптоглобин, и т.д.), секретируемых в основном гепатоцитами – клетками печени. Третья группа - признаки, отражающие уровни N-гликанов белков, секретируемых как B-лимфоцитами и их потомками, так и гепатоцитами. Четвертая группа - признаки, классификация которых не была произведена. Классификация была проведена на основе данных, опубликованных в работе [3], в которой исследователи оценили вклад каждого из N-гликопротеинов в N-гликом плазмы крови человека.

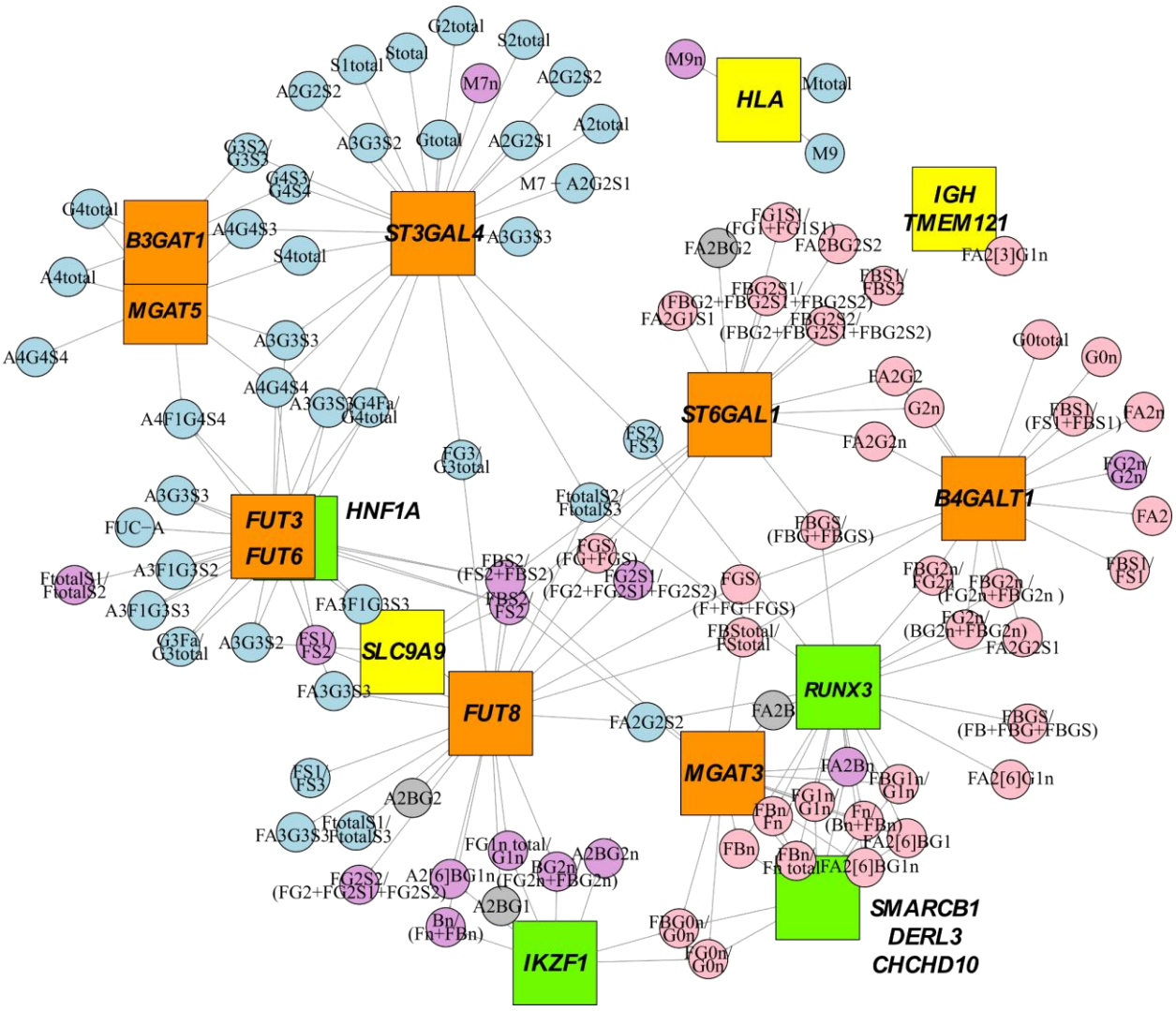


Рис. 17. Генная сеть регуляции уровней N-гликанов белков плазмы крови. Квадратами обозначены локусы, найденные в данном исследовании. Для каждого локуса приведены названия приоритизированных генов. Оранжевым цветом выделены локусы, для которых были приоритизированы гены гликозилтрансфераз. Зеленым цветом окрашены локусы, для которых были приоритизированы гены факторов транскрипции. Кругами обозначены уровни N-гликанов белков плазмы крови. Синим цветом окрашены уровни N-гликанов, связанных с гликопротеинами, секретируемыми гепатоцитами. Розовым цветом окрашены уровни N-гликанов, связанных с гликопротеинами (а именно иммуноглобулинами), секретируемыми клетками лимфоцитарного ряда. Пурпурным цветом окрашены уровни N-гликанов, связанных с как с гликопротеинами, секретируемыми гепатоцитами, так и с гликопротеинами, секретируемыми клетками лимфоцитарного ряда. Серым цветом окрашены уровни N-гликанов, для которых классификация не была проведена. Связи в данной сети обозначают генетическую ассоциацию с P-value $< 2.67 * 10^{-5}$.

Локусы и ассоциированные признаки в данной сети могут быть визуально разделены на две частично перекрывающиеся подсети. Первая подсеть образована локусами *ST3GAL4*, *B3GAT1*, *MGAT5*, *HNF1A*, *FUT3/FUT6*, *FUT8* и *SLC9A9*. Эта подсеть ассоциирована с уровнями N-гликанов, связанных с N-гликопротеинами, секретируемыми в кровоток клетками печени. Большинство из этих признаков

отражают уровни N-гликанов с тремя или четырьмя антеннами, не встречающимися в составе иммуноглобулинов. В данную сеть входит локус *HNF1A* - транскрипционного фактора гепатоцитов. При этом локус *HNF1A* показал ассоциацию с теми же признаками, что и локус *FUT3*/*FUT6*, что соответствует ранее доказанной роли гена *HNF1A* в регуляции уровня экспрессии фукозилтрансфераз [33]. Полученные результаты позволяют выдвинуть гипотезу о том, что данная подсеть генов регулирует процессы гликозилирования в гепатоцитах. Роль генов-кандидатов, входящих в данную сеть, скорее всего следует проверять на клетках печени, например, гепатоцитах, или клетках, близких к ним, например клеточной линии HepG2.

Вторая подсеть образована локусами *FUT8*, *FUT6*/*FUT3*, *SLC9A9*, *IKZF1*, *MGAT3*, *RUNX3*, *SMARCB1*/*DERL3*/*CHCHD10*, *B4GALT1*, *ST6GAL1* и *IGH*/*TMEM121*. Эти локусы ассоциированы с уровнями N-гликанов, связанных с иммуноглобулинами, секретируемыми в кровоток клетками лимфоцитарного ряда. В двух других работах, выполненных в соавторстве с соискателем, было показано, что данные локусы ассоциированы с N-гликозилированием IgG [153, 166]. Основываясь на том, что IgG является наиболее представленным гликопротеином плазмы крови, можно выдвинуть гипотезу о том, что гены-кандидаты из данной сети, регулируют процессы N-гликозилирования в В-лимфоцитах и их потомках. Роль генов-кандидатов, входящих в данную сеть, скорее всего следует проверять на материале клеток, секретирующих антитела или клеток, близких к ним.

Глава 4. Обсуждение

В диссертационной работе описаны результаты первого в мире полногеномного исследования ассоциаций уровней N-гликанов белков плазмы крови человека, измеренных методом СВЭЖХ. В результате данной работы число локусов, ассоциированных с гликомом плазмы крови, было увеличено с 6 до 16. Ассоциация 15 из этих локусов была подтверждена в независимых выборках. Для найденных локусов было проведено биоинформатическое исследование с целью приоритизации генов-кандидатов и установления возможных механизмов влияния данных генов на регуляцию N-гликома плазмы крови. В результате был предложено 18 генов-кандидатов, семь из которых являются новыми - гены регуляторов транскрипции (*IKZF1*, *SMARCB1* и *RUNX3*), деградации гликопротеинов (*DERL3*), тяжелой цепи иммуноглобулинов (*IGH*) и гены с неизвестной функцией (*TMEM121* и *CHCHD10*).

На выборке объемом 2,763 человек была найдена ассоциация 14 локусов, в то время как в предыдущем исследовании на выборке большего объема - 3,533 человека - было найдено 6 локусов [34]. Нахождение большего числа локусов при меньшем размере выборки говорит о более высокой мощности анализа, выполненного в данной диссертационной работе. Как минимум два фактора могли повлиять на увеличение мощности. Технология СВЭЖХ имеет большее разрешение и точность измерения [167] по сравнению с ВЭЖХ, применявшейся ранее в ПГИА N-гликома плазмы крови. Это могло повысить мощность анализа генетических ассоциаций. Показательным для иллюстрации этого утверждения является то, что в настоящем исследовании мы смогли найти и подтвердить ассоциацию двух локусов, содержащих гены *ST3GAL4* и *ST6GAL1*. Эти локусы показали существенно более сильную ассоциацию с уровнями N-гликанов, чем любой из шести ранее найденных локусов, но при этом не были найдены в предыдущих исследованиях [33, 34]. Локус *ST3GAL4* показал наиболее сильную ассоциацию с признаком PGP17 (N-гликан A2G2S[3,6+3]2; P-value ассоциации =

8.6×10^{-28}). Пик PGP17 на хроматограмме расположен рядом с пиками PGP18 и PGP19, средняя площадь под которыми в 8 и 28 раз больше, чем площадь пика PGP17. На хроматограмме ВЭЖХ эти три пика (PGP17, PGP18 и PGP19) объединены в один пик GP9. Таким образом, вклад пика PGP17 в пик GP9 на хроматограмме ВЭЖХ мал по сравнению с пиками PGP18 и PGP19. В данном исследовании локус *ST3GAL4* не показал достоверную ассоциацию с пиками PGP18 (P-value ассоциации = $6.14 * 10^{-1}$) и PGP19 (P-value ассоциации = $8.25 * 10^{-4}$). Таким образом, можно предположить, что ассоциация локуса *ST3GAL4* не была найдена в предыдущих исследования из-за не разделения пиков PGP17, PGP18 и PGP19. Такой же вывод можно сделать относительно ассоциации локуса *ST6GAL1*. В данном исследовании локус показал ассоциацию с признаком PGP13 ($P = 3.12 * 10^{-23}$), который вместе с пиком PGP12 образует пик GP6 на хроматограмме ВЭЖХ. Локус *ST6GAL1* не показал ассоциации ни с пиком GP6 в исследовании [34](P-value ассоциации = 0,012), ни с пиком PGP12 в данном исследовании ($P = 3.19 * 10^{-2}$). Пик PGP12 имеет площадь в 1,5 раза больше, чем пик PGP13. Таким образом, мы предполагаем, что локус *ST6GAL1* ассоциирован исключительно с признаком PGP13 (FA2[3]G1S[3+6]1). Эти примеры убедительно показывают роль улучшенного разрешения разделения N-гликанов методом СВЭЖХ в увеличении мощности анализа генетических ассоциаций.

Другим фактором, который мог увеличить мощность анализа ассоциации, являлось использование геномных данных, имеющих большее разрешение и точность. В предыдущем исследовании генетические данные были импутированы с использованием референтной выборки HapMap2 [44], а в данном исследовании использовались данные современной референтной выборки «1000 Геномов» [134]. Импутация генотипов с использованием данных выборки «1000 Геномов» обеспечивает более высокую точность восстановления неизмеренных генотипов и покрытие генома (8 миллионов ОНП по сравнению с 2.5 миллионами ОНП). Таким образом оба фактора – более современная технология измерения уровней N-

гликанов и генетические данные более высокого разрешения - могли внести вклад в большую мощность анализа генетических ассоциаций.

Среди шести локусов, найденных в предыдущих исследованиях генетического контроля N-гликома плазмы крови [33, 34], только локус *FUT8* был найден в первом ПГИА уровней N-гликанов иммуноглобулина G [64]. В исследовании [166], опубликованном в соавторстве с соискателем, был найден второй локус – *FUT6*/*FUT3*, ассоциированный как с N-гликомом плазмы крови, так и с N-гликомом IgG. Ферменты Fuc-TVI и Fuc-TIII кодируемые данными генами, отвечают за антеннарное фукозилирование. Ранее присутствие антеннарного фукозилирования в структуре иммуноглобулинов не было известно. Можно выдвинуть два предположения о механизме ранее найденной ассоциации локуса *FUT6*/*FUT3* с уровнями N-гликанов IgG [166]: либо ферменты Fuc-TVI и Fuc-TIII проявляют неканоничную активность и переносят остаток фукозы к остову N-гликана, либо антеннарное фукозилирование присутствует в N-гликоме IgG, что было показано в работе [65]. В данной работе не была показана ассоциация локуса *FUT6*/*FUT3* с уровнями N-гликанов с фукозилированием остова, что косвенным образом говорит в пользу второго варианта – существования антеннарного фукозилирования в N-гликоме IgG.

Основными источниками N-гликопротеинов в плазме крови являются клетки печени и клетки, секретирующие антитела [3]. Мы ожидали, что найденные локусы будут обогащены генами, секретирующимися в этих тканях. Действительно, нами было показано, что найденные локусы обогащены генами, экспрессирующимися в плазмоцитах и В-лимфоцитах. Также были найдены эффекты локусов на уровень транскрипции генов-кандидатов в В-лимфоцитах. Однако нами не было обнаружено обогащение генами, экспрессирующимися в печени или гепатоцитах в частности. С одной стороны, это может быть вызвано тем, что только 6 из 15 локусов ассоциированы с уровнями N-гликанов секретируемых печенью (см. раздел 3.5 Генная сеть регуляции N-гликозилирования), в то время как 10 из 15 локусов показали ассоциацию с уровнями N-гликанов иммуноглобулинов. С

другой стороны, мы анализировали данные об экспрессии генов в клетках лимфоцитарного ряда (в частности данные по CD19+ клеткам - B-лимфоцитам - в исследовании CEDAR), в то время как нам не были доступны данные об экспрессии генов в гепатоцитах, а только лишь данные общего транскриптома печени. В будущем для расширения наших знаний о тканеспецифичной регуляции процессов N-гликозилирования потребуется достижение более высокой мощности анализа генетических ассоциаций N-гликома белков плазмы крови, не являющиеся иммуноглобулинами. Это может быть достигнуто как за счет появления новых технологий и протоколов измерения N-гликомных профилей неиммуноглобулиновых белков, так и путем применения более мощных статистических моделей.

Исследование, представленное в рамках диссертационной работы, имеет несколько ограничений. Во-первых, в работе использовались выборки людей, имеющих европейское происхождение. Полученные результаты может быть трудно обобщить на выборки людей с неевропейским происхождением. В то же время, стоит отметить, что в работе [124] было проведено подтверждение результатов ПГИА на выборке QMDiab, образцы которой имеют катарское, филиппинское и южно-азиатское происхождение. Было показано совпадение знака и размера эффекта (в рамках погрешности измерения) локусов, что может говорить об устойчивости результатов ПГИА при анализе выборок неевропейского происхождения. Во-вторых, в диссертационной работе был проведен ПГИА уровней N-гликанов белков плазмы крови без разделения на белок-специфичные профили N-гликозилирования, что уменьшает разрешающую способность анализа тканеспецифичной регуляции N-гликозилирования белков человека. Несмотря на это, реконструкция генной сети регуляции N-гликозилирования позволила сделать несколько выводов о возможных тканеспецифичных подсетях регуляции данного процесса. В-третьих, статистическая мощность подтверждения ассоциации локуса *KREMEN1* равнялась 9%, что оставляет под вопросом достоверность ассоциации локуса гена *KREMEN1* с уровнями N-гликанов белков плазмы крови.

В предыдущих исследованиях были найдены шесть локусов, контролирующих N-гликозилирование белков плазмы крови человека, из которых четыре содержали гены с известной ролью в процессе биосинтеза N-гликанов. В данной диссертационной работе, выполненной с использованием более точной технологии измерения уровней N-гликанов (СВЭЖХ) и генетических данных высокого разрешения, была подтверждена ассоциация шести ранее найденных локусов и найдена и подтверждена ассоциация девяти новых локусов. Совместный анализ гликомных, геномных данных и данных функциональной геномики позволил приоритизировать наиболее вероятные гены-кандидаты в найденных локусах. Реконструкция генной сети регуляции N-гликозилирования позволила выдвинуть гипотезы о тканеспецифичной регуляции N-гликозилирования предложенными генами-кандидатами.

Результаты диссертационной работы подтверждают представление о том, что генетический контроль N-гликозилирования белков плазмы крови человека представляет собой сложный процесс, который находится под контролем генов, принадлежащих разным биологическим путям и экспрессирующимся в разных тканях. На основе полученных результатов сформулированы функциональные гипотезы о возможных механизмах влияния найденных локусов на N-гликозилирование белков плазмы крови. В том числе нами были предложены семь новых генов-кандидатов, вовлеченные в процесс N-гликозилирования белков плазмы крови человека. Эти гипотезы будут востребованы в области молекулярно-генетических исследований гликома и изучения роли гликома в патогенезе социально и экономически важных заболеваний человека.

Существует несколько направлений развития данной работы. Будут проведены более мощные ПГИА уровней N-гликанов. Появятся новые данные функциональной геномики, применимые для изучения процессов N-гликозилирования, совместный анализ которых с результатами ПГИА позволит определить большее число локусов и потенциальных регуляторов N-гликозилирования. На данный момент применение метода ПГИА для изучения

регуляции гликозилирования ограничено анализом общего N-гликома плазмы крови и N-гликома IgG. Развитие технологий профилирования N-гликома расширит набор белков, для которых будут изучены индивидуальные профили N-гликозилирования. С другой стороны, скорее всего появятся высокопроизводительные технологии профилирования N-гликома других тканей человека. Все вышеперечисленное позволит составить более полную картину регуляции N-гликозилирования белков человека, что, в свою очередь, позволит установить роль гликозилирования в патогенезе гликом-ассоциированных заболеваний и ускорить процессы разработки методов прогнозирования, профилактики, диагностики и лечения данных заболеваний.

Заключение

В соответствии с поставленными задачи, были получены следующие результаты:

1. С помощью полногеномного исследования ассоциаций 113 уровней N-гликанов белков плазмы крови на материале 2,763 образцов выборки TwinsUK было обнаружено 14 локусов, для 10 из которых ассоциация с уровнями гликозилирования белков плазмы крови человека была показана впервые. Таким образом, с учетом результатов предыдущих исследований, общее число локусов, ассоциированных с уровнями N-гликанов белков плазмы крови человека было увеличено с 6 до 16.
2. Для того, чтобы проверить полученные результаты на независимой выборке, был разработан и валидирован метод гармонизации данных об уровнях N-гликанов белков плазмы крови человека, измеренных сверхвысокоэффективной жидкостной хроматографией. С помощью разработанного метода были гармонизированы гликомные профили 4,802 участников четырех выборок.
3. На материале четырех независимых выборок была подтверждена ассоциация 15 из 16 локусов с уровнями N-гликанов белков плазмы крови, в том числе 9 локусов, ассоциация которых впервые была показана в данном исследовании.
4. Для 15 локусов с подтверждённой ассоциацией был проведен ряд анализов с использованием количественно-генетических и биоинформатических методов, позволивший предложить 18 генов-кандидатов, наиболее вероятно влияющий на уровни N-гликанов белков плазмы крови человека.

Выводы

1. В результате проведения полногеномного исследования ассоциаций генетических маркеров с уровнями N-гликанов белков плазмы крови человека и последующего подтверждения найденных ассоциаций, показано участие 15 локусов генома в контроле исследованных признаков. В данных локусах приоритизированы 18 генов-кандидатов, чьи продукты наиболее вероятно вовлечены в регуляцию N-гликозилирования белков плазмы крови человека.

2. Для девяти локусов подтвержденная ассоциация с уровнями N-гликанов белков плазмы крови была показана впервые. Для этих локусов были предложены гены-кандидаты *IKZF1*, *RUNX3*, *SMARCB1*, *DERL3*, *CHCHD10*, *IGH* и *TMEM121*, чья роль в регуляции N-гликозилирования белков плазмы крови человека ранее не была известна.

3. Полученные результаты подтверждают вовлеченность генов метаболизма N-гликанов и их факторов транскрипции в регуляцию N-гликозилирования белков плазмы крови человека. Кроме того, результаты позволяют предположить участие в регуляции данного процесса генов деградации гликопротеинов и гомеостаза pH в аппарате Гольджи.

4. Реконструкция генной сети позволила предположить, что часть приоритизированных нами генов регулирует N-гликозилирование белков в печени, а часть - в клетках, секретирующих антитела.

Список использованной литературы

1. Khoury G.A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database / Khoury G.A., Baliban R.C., Floudas C.A. // Scientific Reports – 2011. – Т. 1 – № 1 – С.90.

2. Craveur P. PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins / Craveur P., Rebehmed J., Brevern A.G. de // Database – 2014. – Т. 2014 – С.bau041.

3. Clerc F. Human plasma protein N-glycosylation / Clerc F., Reiding K.R., Jansen B.C., Kammeijer G.S.M., Bondt A., Wuhrer M. // Glycoconjugate Journal – 2016. – Т. 33 – № 3 – С.309–343.

4. Varki A. Biological roles of oligosaccharides: all of the theories are correct / Varki A. // Glycobiology – 1993. – Т. 3 – № 2 – С.97–130.

5. Ohtsubo K. Glycosylation in Cellular Mechanisms of Health and Disease / Ohtsubo K., Marth J.D. // Cell – 2006. – Т. 126 – № 5 – С.855–867.

6. Skropeta D. The effect of individual N-glycans on enzyme activity / Skropeta D. // Bioorganic & Medicinal Chemistry – 2009. – Т. 17 – № 7 – С.2645–2653.

7. Varki A. Biological roles of glycans / Varki A. // Glycobiology – 2017. – Т. 27 – № 1 – С.3–49.

8. Lauc G. Mechanisms of disease: The human N-glycome. / Lauc G., Pezer M., Rudan I., Campbell H. // Biochimica et biophysica acta – 2015. – Т. 1860 – № 8 – С.1574–1582.

9. Poole J. Glycointeractions in bacterial pathogenesis / Poole J., Day C.J., Itzstein M. von, Paton J.C., Jennings M.P. // Nature Reviews Microbiology – 2018. – Т. 16 – № 7 – С.440–452.

10. Chang I.J. Congenital disorders of glycosylation / Chang I.J., He M., Lam C.T.

// Annals of Translational Medicine – 2018. – Т. 6 – № 24 – С.477–477.

11. Freeze H.H. Genetic defects in the human glycome / Freeze H.H. // Nature Reviews Genetics – 2006. – Т. 7 – № 7 – С.537–551.

12. Reily C. Glycosylation in health and disease / Reily C., Stewart T.J., Renfrow M.B., Novak J. // Nature Reviews Nephrology – 2019. – Т. 15 – № 6 – С.346–366.

13. Fuster M.M. The sweet and sour of cancer: glycans as novel therapeutic targets / Fuster M.M., Esko J.D. // Nature Reviews Cancer – 2005. – Т. 5 – № 7 – С.526–542.

14. Connelly M.A. Inflammatory glycoproteins in cardiometabolic disorders, autoimmune diseases and cancer / Connelly M.A., Gruppen E.G., Otvos J.D., Dullaart R.P.F. // Clinica Chimica Acta – 2016. – Т. 459 – С.177–186.

15. Dube D.H. Glycans in cancer and inflammation — potential for therapeutics and diagnostics / Dube D.H., Bertozzi C.R. // Nature Reviews Drug Discovery – 2005. – Т. 4 – № 6 – С.477–488.

16. Pagan J.D. Engineered Sialylation of Pathogenic Antibodies In Vivo Attenuates Autoimmune Disease / Pagan J.D., Kitaoka M., Anthony R.M. // Cell – 2018. – Т. 172 – № 3 – С.564–577.

17. Adamczyk B. Glycans as cancer biomarkers / Adamczyk B., Tharmalingam T., Rudd P.M. // Biochimica et Biophysica Acta (BBA) - General Subjects – 2012. – Т. 1820 – № 9 – С.1347–1353.

18. Maverakis E. Glycans in the immune system and The Altered Glycan Theory of Autoimmunity: A critical review / Maverakis E., Kim K., Shimoda M., Gershwin M.E., Patel F., Wilken R., Raychaudhuri S., Ruhaak L.R., Lebrilla C.B. // Journal of Autoimmunity – 2015. – Т. 57 – С.1–13.

19. RodrÍguez E. The tumour glyco-code as a novel immune checkpoint for immunotherapy. / RodrÍguez E., Schetters S.T.T., Kooyk Y. van // Nature reviews. Immunology – 2018. – Т. 18 – № 3 – С.204–211.

20. Thanabalasingham G. Mutations in HNF1A result in marked alterations of plasma glycan profile / Thanabalasingham G., Huffman J.E., Kattla J.J., Novokmet M., Rudan I., Gloyn A.L., Hayward C., Adamczyk B., Reynolds R.M., Muzinic A., Hassanali N., Pucic M., Bennett A.J., Essafi A., Polasek O., Mughal S.A., Redzic I., Primorac D., Zgaga L., Kolcic I., Hansen T., Gasperikova D., Tjora E., Strachan M.W.J., Nielsen T., Stanik J., Klimes I., Pedersen O.B., Njølstad P.R., Wild S.H., Gyllensten U., Gornik O., Wilson J.F., Hastie N.D., Campbell H., McCarthy M.I., Rudd P.M., Owen K.R., Lauc G., Wright A.F. // Diabetes – 2013. – Т. 62 – № 4 – С.1329–1337.

21. Gudelj I. Low galactosylation of IgG associates with higher risk for future diagnosis of rheumatoid arthritis during 10 years of follow-up / Gudelj I., Salo P.P., Trbojević-Akmačić I., Albers M., Primorac D., Perola M., Lauc G. // Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease – 2018. – Т. 1864 – № 6 – С.2034–2039.

22. Visscher P.M. Five Years of GWAS Discovery / Visscher P.M., Brown M.A., McCarthy M.I., Yang J. // The American Journal of Human Genetics – 2012. – Т. 90 – № 1 – С.7–24.

23. Visscher P.M. 10 Years of GWAS Discovery: Biology, Function, and Translation / Visscher P.M., Wray N.R., Zhang Q., Sklar P., McCarthy M.I., Brown M.A., Yang J. // The American Journal of Human Genetics – 2017. – Т. 101 – № 1 – С.5–22.

24. Varki A. Nothing in Glycobiology Makes Sense, except in the Light of Evolution / Varki A. // Cell – 2006. – Т. 126 – № 5 – С.841–845.

25. Corfield A.P. Glycan variation and evolution in the eukaryotes / Corfield A.P., Berry M. // Trends in Biochemical Sciences – 2015. – Т. 40 – № 7 – С.351–359.

26. Lombard V. The carbohydrate-active enzymes database (CAZy) in 2013 / Lombard V., Golaconda Ramulu H., Drula E., Coutinho P.M., Henrissat B. // Nucleic Acids Research – 2014. – Т. 42 – № D1 – С.D490–D495.

27. Kukuruzinska M.A. Protein N-Glycosylation: Molecular Genetics and

Functional Significance / Kukuruzinska M.A., Lennon K. // Critical Reviews in Oral Biology & Medicine – 1998. – Т. 9 – № 4 – С.415–448.

28. Nairn A. V Regulation of glycan structures in animal tissues: transcript profiling of glycan-related genes / Nairn A. V, York W.S., Harris K., Hall E.M., Pierce J.M., Moremen K.W. // J. Biol. Chem. – 2008. – Т. 283 – № 25 – С.17298–17313.

29. Nairn A. V Regulation of glycan structures in murine embryonic stem cells: combined transcript profiling of glycan-related genes and glycan structural analysis / Nairn A. V, Aoki K., Rosa M. dela, Porterfield M., Lim J.-M., Kulik M., Pierce J.M., Wells L., Dalton S., Tiemeyer M., Moremen K.W. // J. Biol. Chem. – 2012. – Т. 287 – № 45 – С.37835–37856.

30. Moremen K.W. Vertebrate protein glycosylation: diversity, synthesis and function / Moremen K.W., Tiemeyer M., Nairn A. V. // Nature Reviews Molecular Cell Biology – 2012. – Т. 13 – № 7 – С.448–462.

31. Stanley P. N-Glycans / под ред. V. Ajit, C. Richard D, E. Jeffrey D, F. Hudson H, S. Pamela, B. Carolyn R, H. Gerald W, E. Marilynn E. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2009. Вып. 2.

32. Lauc G. Complex genetic regulation of proteinglycosylation / Lauc G., Rudan I., Campbell H., Rudd P.M. // Mol. BioSyst. – 2010. – Т. 6 – № 2 – С.329–335.

33. Lauc G. Genomics Meets Glycomics—The First GWAS Study of Human N-Glycome Identifies HNF1α as a Master Regulator of Plasma Protein Fucosylation / Lauc G., Essafi A., Huffman J.E., Hayward C., Knežević A., Kattla J.J., Polašek O., Gornik O., Vitart V., Abrahams J.L., Pučić M., Novokmet M., Redžić I., Campbell S., Wild S.H., Borovečki F., Wang W., Kolčić I., Zgaga L., Gyllensten U., Wilson J.F., Wright A.F., Hastie N.D., Campbell H., Rudd P.M., Rudan I. // PLoS Genetics – 2010. – Т. 6 – № 12 – С.e1001256.

34. Huffman J.E. Polymorphisms in B3GAT1, SLC9A9 and MGAT5 are associated with variation within the human plasma N-glycome of 3533 European adults /

Huffman J.E., Knežević A., Vitart V., Kattla J., Adamczyk B., Novokmet M., Igl W., Pučić M., Zgaga L., Johannson Å., Redžić I., Gornik O., Zemunik T., Polašek O., Kolčić I., Pehlić M., Koeleman C.A.M., Campbell S., Wild S.H., Hastie N.D., Campbell H., Gyllensten U., Wuhrer M., Wilson J.F., Hayward C., Rudan I., Rudd P.M., Wright A.F., Lauc G. // Human Molecular Genetics – 2011. – T. 20 – № 24 – C.5000–5011.

35. Yamagata K. Mutations in the hepatocyte nuclear factor-1α gene in maturity-onset diabetes of the young (MODY3) / Yamagata K., Oda N., Kaisaki P.J., Menzel S., Furuta H., Vaxillaire M., Southam L., Cox R.D., Lathrop G.M., Boriraj V.V., Chen X., Cox N.J., Oda Y., Yano H., Beau M.M. Le, Yamada S., Nishigori H., Takeda J., Fajans S.S., Hattersley A.T., Iwasaki N., Hansen T., Pedersen O., Polonsky K.S., Turner R.C., Velho G., Chèvre J.-C., Froguel P., Bell G.I. // Nature – 1996. – T. 384 – № 6608 – C.455–458.

36. Knežević A. High throughput plasma N-glycome profiling using multiplexed labelling and UPLC with fluorescence detection / Knežević A., Bones J., Kračun S.K., Gornik O., Rudd P.M., Lauc G. // The Analyst – 2011. – T. 136 – № 22 – C.4670–4673.

37. Agakova A. Automated Integration of a UPLC Glycomic Profile / Agakova A., Vučković F., Klarić L., Lauc G., Agakov F. // Methods in Molecular Biology – 2017. – T. 1503 – C.217–233.

38. Huffman J.E. Comparative Performance of Four Methods for High-throughput Glycosylation Analysis of Immunoglobulin G in Genetic and Epidemiological Research / Huffman J.E., Pučić-Baković M., Klarić L., Hennig R., Selman M.H.J., Vučković F., Novokmet M., Krištić J., Borowiak M., Muth T., Polašek O., Razdorov G., Gornik O., Plomp R., Theodoratou E., Wright A.F., Rudan I., Hayward C., Campbell H., Deelder A.M., Reichl U., Aulchenko Y.S., Rapp E., Wuhrer M., Lauc G. // Molecular & Cellular Proteomics – 2014. – T. 13 – № 6 – C.1598–1610.

39. Trbojević Akmačić I. High-throughput glycomics: Optimization of sample preparation / Trbojević Akmačić I., Ugrina I., Štambuk J., Gudelj I., Vučković F., Lauc G., Pučić-Baković M. // Biochemistry (Moscow) – 2015. – T. 80 – № 7 – C.934–942.

40. Durbin R.M. A map of human genome variation from population-scale sequencing / Durbin R.M., Altshuler D.L., Durbin R.M., McVean G.A. // Nature – 2010. – T. 467 – № 7319 – C.1061–1073.

41. McCarthy S. A reference panel of 64,976 haplotypes for genotype imputation / McCarthy S., Das S., Kretzschmar W., Marchini J. // Nature Genetics – 2016. – T. 48 – № 10 – C.1279–1283.

42. Taliun D. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program / Taliun D., Harris D.N., Kessler M.D., Abecasis G.R. // Nature – 2021. – T. 590 – № 7845 – C.290–299.

43. Consortium I.H. The International HapMap Project / Consortium I.H. // Nature – 2003. – T. 426 – № 6968 – C.789–796.

44. Frazer K.A. A second generation human haplotype map of over 3.1 million SNPs / Frazer K.A., Ballinger D.G., Cox D.R., Stewart J. // Nature – 2007. – T. 449 – № 7164 – C.851–861.

45. Saldova R. Association of N-glycosylation with breast carcinoma and systemic features using high-resolution quantitative UPLC / Saldova R., Asadi Shehni A., Haakensen V.D., Steinfeld I., Hilliard M., Kifer I., Helland Å., Yakhini Z., Børresen-Dale A.-L.L., Rudd P.M., Helland A., Yakhini Z., Børresen-Dale A.-L.L., Rudd P.M. // Journal of Proteome Research – 2014. – T. 13 – № 5 – C.2314–2327.

46. Gudelj I. Estimation of human age using N-glycan profiles from bloodstains / Gudelj I., Keser T., Vučković F., Škaro V., Goreta S.Š., Pavić T., Dumić J., Primorac D., Lauc G., Gornik O. // International Journal of Legal Medicine – 2015. – T. 129 – № 5 – C.955–961.

47. Gudelj I. Changes in total plasma and serum N-glycome composition and patient-controlled analgesia after major abdominal surgery / Gudelj I., Baciarello M., Ugrina I., Allegri M. // Scientific Reports – 2016. – T. 6 – C.31234.

48. Anderson C.A. Data quality control in genetic case-control association studies

/ Anderson C.A., Pettersson F.H., Clarke G.M., Cardon L.R., Morris A.P., Zondervan K.T. // Nature Protocols – 2010. – Т. 5 – № 9 – С.1564–1573.

49. Winkler T.W. Quality control and conduct of genome-wide association meta-analyses / Winkler T.W., Day F.R., Croteau-Chonka D.C., Wood A.R., Locke A.E., Mägi R., Ferreira T., Fall T., Graff M., Justice A.E., Luan J., Gustafsson S., Randall J.C., Vedantam S., Workalemahu T., Kilpeläinen T.O., Scherag A., Esko T., Kutalik Z., Heid I.M., Loos R.J.F., Consortium T.G.I. of A.T. (GIANT) // Nature Protocols – 2014. – Т. 9 – № 5 – С.1192–1212.

50. GTEx Consortium Genetic effects on gene expression across human tissues. / GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)---Analysis Working Group, Statistical Methods groups---Analysis Working Group, Montgomery S.B. // Nature – 2017. – Т. 550 – № 7675 – С.204–213.

51. Momozawa Y. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes / Momozawa Y., Dmitrieva J., Théâtre E., Deffontaine V., Rahmouni S., Charloteaux B., Crins F., Docampo E., Elansary M., Gori A.-S., Lecut C., Mariman R., Mni M., Oury C., Altukhov I., Alexeev D., Aulchenko Y., Amininejad L., Bouma G., Hoentjen F., Löwenberg M., Oldenburg B., Pierik M.J., Meulen-de Jong A.E. vander, Janneke van der Woude C., Visschedijk M.C., Lathrop M., Hugot J.-P., Weersma R.K., Vos M. De, Franchimont D., Vermeire S., Kubo M., Louis E., Georges M. // Nature Communications – 2018. – Т. 9 – № 1 – С.2427.

52. McLaren W. The Ensembl Variant Effect Predictor / McLaren W., Gil L., Hunt S.E., Riat H.S., Ritchie G.R.S., Thormann A., Flicek P., Cunningham F. // Genome Biology – 2016. – Т. 17 – № 1 – С.122.

53. Rogers M.F. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features / Rogers M.F., Shihab H.A., Mort M., Cooper D.N., Gaunt T.R., Campbell C. // Bioinformatics – 2017. – Т. 34 – № 3 – С.511–513.

54. Ferlaino M. An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome / Ferlaino M., Rogers M.F.,

Shihab H.A., Mort M., Cooper D.N., Gaunt T.R., Campbell C. // BMC Bioinformatics – 2017. – Т. 18 – № 1 – С.442.

55. Zhu Z. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets / Zhu Z., Zhang F., Hu H., Bakshi A., Robinson M.R., Powell J.E., Montgomery G.W., Goddard M.E., Wray N.R., Visscher P.M., Yang J. // Nature Genetics – 2016. – Т. 48 – № 5 – С.481–487.

56. Pers T.H. Biological interpretation of genome-wide association studies using predicted gene functions / Pers T.H., Karjalainen J.M., Chan Y., Westra H.-J., Wood A.R., Yang J., Lui J.C., Vedantam S., Gustafsson S., Esko T., Frayling T., Speliotes E.K., Boehnke M., Raychaudhuri S., Fehrmann R.S.N., Hirschhorn J.N., Franke L. // Nature Communications – 2015. – Т. 6 – № 1 – С.5890.

57. Varki A. Historical Background and Overview / под ред. A. Varki, R.D. Cummings, J.D. Esko, H.H. Freeze, P. Stanley, C.R. Bertozzi, G.W. Hart, M.E. Etzler. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2015. Вып. 3.

58. Gagneux P. Evolution of Glycan Diversity / под ред. A. Varki, R.D. Cummings, J.D. Esko, H.H. Freeze, P. Stanley, C.R. Bertozzi, G.W. Hart, M.E. Etzler. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2015. Вып. 3.

59. Lauc G. Glycans - the third revolution in evolution. / Lauc G., Krištić J., Zoldoš V. // Frontiers in genetics – 2014. – Т. 5 – С.145.

60. Adamczyk B. Pregnancy-Associated Changes of IgG and Serum N-Glycosylation in Camel (Camelus dromedarius) / Adamczyk B., Albrecht S., Stöckmann H., Ghoneim I.M., Al-Eknah M., Al-Busadah K.A.S., Karlsson N.G., Carrington S.D., Rudd P.M. // Journal of Proteome Research – 2016. – Т. 15 – № 9 – С.3255–3265.

61. Krištić J. Glycans are a novel biomarker of chronological and biological ages / Krištić J., Vučković F., Menni C., Klarić L., Keser T., Beceheli I., Pučić-Baković M., Novokmet M., Mangino M., Thaqi K., Rudan P., Novokmet N., Šarac J., Missoni S., Kolčić I., Polašek O., Rudan I., Campbell H., Hayward C., Aulchenko Y., Valdes A.,

Wilson J.F., Gornik O., Primorac D., Zoldoš V., Spector T., Lauc G., Sarac J., Missoni S., Kolčić I., Polašek O., Rudan I., Campbell H., Hayward C., Aulchenko Y., Valdes A., Wilson J.F., Gornik O., Primorac D., Zoldoš V., Spector T., Lauc G. // Journals of Gerontology - Series A Biological Sciences and Medical Sciences – 2014. – Т. 69 – № 7 – С.779–789.

62. Anthony R.M. Novel roles for the IgG Fc glycan / Anthony R.M., Wermeling F., Ravetch J. V. // Annals of the New York Academy of Sciences – 2012. – Т. 1253 – № 1 – С.170–180.

63. Freidin M.B. The Association Between Low Back Pain and Composition of IgG Glycome. / Freidin M.B., Keser T., Gudelj I., Štambuk J., Vučenović D., Allegri M., Pavić T., Šimurina M., Fabiane S.M., Lauc G., Williams F.M.K. // Scientific reports – 2016. – Т. 6 – С.26815.

64. Lauc G. Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. / Lauc G., Huffman J.E., Pučić M., Zgaga L., Adamczyk B., Mužinić A., Novokmet M., Polašek O., Gornik O., Krištić J., Keser T., Vitart V., Scheijen B., Uh H.-W.W., Molokhia M., Patrick A.L., McKeigue P., Kolčić I., Lukić I.K., Swann O., Leeuwen F.N. van, Ruhaak L.R., Houwing-Duistermaat J.J., Slagboom P.E., Beekman M., Craen A.J.M.M. de, Deelder A.M., Zeng Q., Wang W., Hastie N.D., Gyllensten U., Wilson J.F., Wuhrer M., Wright A.F., Rudd P.M., Hayward C., Aulchenko Y., Campbell H., Rudan I. // PLoS genetics – 2013. – Т. 9 – № 1 – С.e1003225.

65. Russell A.C. The N-glycosylation of immunoglobulin G as a novel biomarker of Parkinson’s disease. / Russell A.C., Šimurina M., Garcia M.T., Novokmet M., Wang Y., Rudan I., Campbell H., Lauc G., Thomas M.G., Wang W. // Glycobiology – 2017. – Т. 27 – № 5 – С.501–510.

66. Lemmers R.F.H. IgG glycan patterns are associated with type 2 diabetes in independent European populations / Lemmers R.F.H., Vilaj M., Urda D., Agakov F., Šimurina M., Klaric L., Rudan I., Campbell H., Hayward C., Wilson J.F., Lieverse A.G.,

Gornik O., Sijbrands E.J.G., Lauc G., Hoek M. van // Biochimica et Biophysica Acta (BBA) - General Subjects – 2017. – Т. 1861 – № 9 – С.2240–2249.

67. Gudelj I. Immunoglobulin G glycosylation in aging and diseases / Gudelj I., Lauc G., Pezer M. // Cellular Immunology – 2018. – Т. 333 – С.65–79.

68. Dotz V. N -glycome signatures in human plasma: associations with physiology and major diseases / Dotz V., Wuhrer M. // FEBS Letters – 2019. – Т. 593 – № 21 – С.2966–2976.

69. Keser T. Increased plasma N-glycome complexity is associated with higher risk of type 2 diabetes / Keser T., Gornik I., Vučković F., Selak N., Pavić T., Lukić E., Gudelj I., Gašparović H., Biočina B., Tilin T., Wennerström A., Männistö S., Salomaa V., Havulinna A., Wang W., Wilson J.F., Charutvedi N., Perola M., Campbell H., Lauc G., Gornik O., Hr O. // Diabetologia – 2017. – Т. 60 – № 12 – С.1–9.

70. Miura Y. Glycomics and glycoproteomics focused on aging and age-related diseases — Glycans as a potential biomarker for physiological alterations / Miura Y., Endo T. // Biochimica et Biophysica Acta (BBA) - General Subjects – 2016. – Т. 1860 – № 8 – С.1608–1614.

71. Trbojević Akmačić I. Inflammatory Bowel Disease Associates with Proinflammatory Potential of the Immunoglobulin G Glycome / Trbojević Akmačić I., Ventham N.T., Theodoratou E., Vučković F., Kennedy N.A., Krištić J., Nimmo E.R., Kalla R., Drummond H., Štambuk J., Dunlop M.G., Novokmet M., Aulchenko Y., Gornik O., Campbell H., Pučić Baković M., Satsangi J., Lauc G. // Inflammatory Bowel Diseases – 2015. – Т. 21 – № 6 – С.1237–1247.

72. Clerc F. Plasma N-Glycan Signatures Are Associated With Features of Inflammatory Bowel Diseases / Clerc F., Novokmet M., Dotz V., Heuvel T. van den // Gastroenterology – 2018. – Т. 155 – № 3 – С.829–843.

73. Wang Y. The Association Between Glycosylation of Immunoglobulin G and Hypertension / Wang Y., Klarić L., Yu X., Thaqi K., Dong J., Novokmet M., Wilson J.,

Polasek O., Liu Y., Krištić J., Ge S., Pučić-Baković M., Wu L., Zhou Y., Ugrina I., Song M., Zhang J., Guo X., Zeng Q., Rudan I., Campbell H., Aulchenko Y., Lauc G., Wang W. // Medicine – 2016. – Т. 95 – № 17 – С.e3379.

74. Taniguchi N. Glycans and cancer: role of N-glycans in cancer biomarker, progression and metastasis, and therapeutics. / Taniguchi N., Kizuka Y. // Advances in cancer research – 2015. – Т. 126 – С.11–51.

75. Mehta A. Glycosylation and liver cancer. / Mehta A., Herrera H., Block T. // Advances in cancer research – 2015. – Т. 126 – С.257–279.

76. Yu X. Profiling IgG N-glycans as potential biomarker of chronological and biological ages / Yu X., Wang Y., Kristic J., Dong J., Chu X., Ge S., Wang H., Fang H., Gao Q., Liu D., Zhao Z., Peng H., Pucic Bakovic M., Wu L., Song M., Rudan I., Campbell H., Lauc G., Wang W. // Medicine – 2016. – Т. 95 – № 28 – С.e4112.

77. Cobb B.A. The history of IgG glycosylation and where we are now / Cobb B.A. // Glycobiology – 2020. – Т. 30 – № 4 – С.202–213.

78. Peipp M. Antibody fucosylation differentially impacts cytotoxicity mediated by NK and PMN effector cells / Peipp M., Lammerts van Bueren J.J., Schneider-Merck T., Bleeker W.W.K., Dechant M., Beyer T., Repp R., Berkel P.H.C. van, Vink T., Winkel J.G.J. van de, Parren P.W.H.I., Valerius T. // Blood – 2008. – Т. 112 – № 6 – С.2390–2399.

79. Mizushima T. Structural basis for improved efficacy of therapeutic antibodies on defucosylation of their Fc glycans / Mizushima T., Yagi H., Takemoto E., Shibata-Koyama M., Isoda Y., Iida S., Masuda K., Satoh M., Kato K. // Genes to Cells – 2011. – Т. 16 – № 11 – С.1071–1080.

80. Chauhan J.S. In silico Platform for Prediction of N-, O- and C-Glycosites in Eukaryotic Protein Sequences / Chauhan J.S., Rao A., Raghava G.P.S. // PLoS ONE – 2013. – Т. 8 – № 6 – С.e67008.

81. Vilaj M. Evaluation of different PNGase F enzymes in immunoglobulin G and

total plasma N-glycans analysis / Vilaj M., Lauc G., Trbojević-Akmačić I. // Glycobiology – 2020. – Т. 31 – № 1 – С.2–7.

82. Mulloy B. Structural Analysis of Glycans / под ред. A. Varki, R.D. Cummings, J.D. Esko, H.H. Freeze, P. Stanley, C.R. Bertozzi, G.W. Hart, M.E. Etzler. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 2015.

83. Harvey D.J. Proposal for a standard system for drawing structural diagrams of N - and O -linked carbohydrates and related compounds / Harvey D.J., Merry A.H., Royle L., P. Campbell M., Dwek R.A., Rudd P.M. // PROTEOMICS – 2009. – Т. 9 – № 15 – С.3796–3801.

84. Nairn A. V. Regulation of Glycan Structures in Animal Tissues / Nairn A. V., York W.S., Harris K., Hall E.M., Pierce J.M., Moremen K.W. // Journal of Biological Chemistry – 2008. – Т. 283 – № 25 – С.17298–17313.

85. Kanehisa M. KEGG: new perspectives on genomes, pathways, diseases and drugs / Kanehisa M., Furumichi M., Tanabe M., Sato Y., Morishima K. // Nucleic Acids Research – 2017. – Т. 45 – № D1 – С.D353–D361.

86. Krištić J. Quantitative Genetics of Human Protein N-Glycosylation. / Krištić J., Sharapov S.Z., Aulchenko Y.S. // Advances in experimental medicine and biology – 2021. – Т. 1325 – С.151–171.

87. Adua E. Innovation Analysis on Postgenomic Biomarkers: Glycomics for Chronic Diseases / Adua E., Russell A., Roberts P., Wang Y., Song M., Wang W. // OMICS: A Journal of Integrative Biology – 2017. – Т. 21 – № 4 – С.183–196.

88. Dagostino C. Validation of standard operating procedures in a multicenter retrospective study to identify -omics biomarkers for chronic low back pain / Dagostino C., Gregori M. De, Gieger C., Manz J., Gudelj I., Lauc G., Divizia L., Wang W., Sim M., Pemberton I.K., MacDougall J., Williams F., Zundert J. Van, Primorac D., Aulchenko Y., Kapural L., Allegri M., PainOmics Group // PLOS ONE – 2017. – Т. 12 – № 5 – С.e0176372.

89. Uhlén M. The human secretome / Uhlén M., Karlsson M.J., Hober A., Sivertsson Å. // Science Signaling – 2019. – Т. 12 – № 609 – С.eaaz0274.

90. Knezevic A. Variability, Heritability and Environmental Determinants of Human Plasma N-Glycome / Knezevic A., Polasek O., Gornik O., Rudan I., Campbell H., Hayward C., Wright A., Kolcic I., O'Donoghue N., Bones J., Others, Knežević A., Polašek O., Gornik O., Rudan I., Campbell H., Hayward C., Wright A., Kolčić I., O'Donoghue N., Bones J., Rudd P.M., Lauc G., O'Donoghue N., Bones J., Rudd P.M., Lauc G. // Journal of Proteome Research – 2009. – Т. 8 – № 2 – С.694–701.

91. Zaytseva O.O. Heritability of Human Plasma N-Glycome / Zaytseva O.O., Freidin M.B., Keser T., Štambuk J., Ugrina I., Šimurina M., Vilaj M., Štambuk T., Trbojević-Akmačić I., Pučić-Baković M., Lauc G., Williams F.M.K.K., Novokmet M. // Journal of Proteome Research – 2020. – Т. 19 – № 1 – С.85–91.

92. Аксенович Т.И.Картирование генов с помощью неравновесия по сцеплению или аллельных ассоциаций: учеб. пособие / Т. И. Аксенович, Н. М. Белоногова – НГУ, 2008.– 98с.

93. Marchini J. Genotype imputation for genome-wide association studies / Marchini J., Howie B. // Nature Reviews Genetics – 2010. – Т. 11 – № 7 – С.499–511.

94. Pasaniuc B. Dissecting the genetics of complex traits using summary association statistics / Pasaniuc B., Price A.L. // Nature Reviews Genetics – 2016. – Т. 18 – № 2 – С.117–127.

95. Klein R.J. Complement Factor H Polymorphism in Age-Related Macular Degeneration / Klein R.J., Zeiss C., Chew E.Y., Tsai J.-Y., Sackler R.S., Haynes C., Henning A.K., SanGiovanni J.P., Mane S.M., Mayne S.T., Bracken M.B., Ferris F.L., Ott J., Barnstable C., Hoh J. // Science – 2005. – Т. 308 – № 5720 – С.385–389.

96. Yu W. GWAS Integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies / Yu W., Yesupriya A., Wulf A., Hindorff L.A., Dowling N., Khoury M.J., Gwinn M. // European Journal of

Human Genetics – 2011. – Т. 19 – № 10 – С.1095–1099.

97. Li M. An imputation approach for oligonucleotide microarrays. / Li M., Wen Y., Lu Q., Fu W.J. // PloS one – 2013. – Т. 8 – № 3 – С.e58677.

98. Martin A.R. Imputation-based assessment of next generation rare exome variant arrays. / Martin A.R., Tse G., Bustamante C.D., Kenny E.E. // Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing – 2014. – Т. 19 – С.241–252.

99. Shi S. Comprehensive Assessment of Genotype Imputation Performance / Shi S., Yuan N., Yang M., Du Z., Wang J., Sheng X., Wu J., Xiao J. // Human Heredity – 2018. – Т. 83 – № 3 – С.107–116.

100. Clarke G.M. Basic statistical analysis in genetic case-control studies / Clarke G.M., Anderson C.A., Pettersson F.H., Cardon L.R., Morris A.P., Zondervan K.T. // Nature Protocols – 2011. – Т. 6 – № 2 – С.121–133.

101. Yang J. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits / Yang J., Ferreira T., Morris A.P., Medland S.E., Madden P.A.F., Heath A.C., Martin N.G., Montgomery G.W., Weedon M.N., Loos R.J., Frayling T.M., McCarthy M.I., Hirschhorn J.N., Goddard M.E., Visscher P.M. // Nature Genetics – 2012. – Т. 44 – № 4 – С.369–375.

102. Hemani G. The MR-Base platform supports systematic causal inference across the human phenome / Hemani G., Zheng J., Elsworth B., Wade K.H., Haberland V., Baird D., Laurin C., Burgess S., Bowden J., Langdon R., Tan V.Y., Yarmolinsky J., Shihab H.A., Timpson N.J., Evans D.M., Relton C., Martin R.M., Davey Smith G., Gaunt T.R., Haycock P.C. // eLife – 2018. – Т. 7.

103. Bulik-Sullivan B. An atlas of genetic correlations across human diseases and traits / Bulik-Sullivan B., Finucane H.K., Anttila V., Gusev A., Day F.R., Loh P.-R., Duncan L., Perry J.R.B., Patterson N., Robinson E.B., Daly M.J., Price A.L., Neale B.M. // Nature Genetics – 2015. – Т. 47 – № 11 – С.1236–1241.

104. Staley J.R. PhenoScanner: a database of human genotype–phenotype

associations / Staley J.R., Blackshaw J., Kamat M.A., Ellis S., Surendran P., Sun B.B., Paul D.S., Freitag D., Burgess S., Danesh J., Young R., Butterworth A.S. // Bioinformatics – 2016. – Т. 32 – № 20 – С.3207–3209.

105. Westra H.-J. Systematic identification of trans eQTLs as putative drivers of known disease associations. / Westra H.-J., Peters M.J., Esko T., Franke L. // Nature genetics – 2013. – Т. 45 – № 10 – С.1238–1243.

106. GTEx Consortium Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. / GTEx Consortium // Science (New York, N.Y.) – 2015. – Т. 348 – № 6235 – С.648–660.

107. Rivinoja A. Elevated Golgi pH impairs terminal *N* -glycosylation by inducing mislocalization of Golgi glycosyltransferases / Rivinoja A., Hassinen A., Kokkonen N., Kauppila A., Kellokumpu S. // Journal of Cellular Physiology – 2009. – Т. 220 – № 1 – С.144–154.

108. Ellard S. Hepatocyte nuclear factor 1 alpha (HNF-1?) mutations in maturity-onset diabetes of the young / Ellard S. // Human Mutation – 2000. – Т. 16 – № 5 – С.377–385.

109. Moayyeri A. The UK Adult Twin Registry (TwinsUK Resource) / Moayyeri A., Hammond C.J., Hart D.J., Spector T.D. // Twin Research and Human Genetics – 2013. – Т. 16 – № 01 – С.144–149.

110. Spector T.D. The UK Adult Twin Registry (TwinsUK). / Spector T.D., Williams F.M.K. // Twin research and human genetics : the official journal of the International Society for Twin Studies – 2006. – Т. 9 – № 6 – С.899–906.

111. Boeing H. Recruitment Procedures of EPIC-Germany / Boeing H., Korfmann A., Bergmann M.M. // Annals of Nutrition and Metabolism – 1999. – Т. 43 – № 4 – С.205–215.

112. Allegri M. ‘Omics’ biomarkers associated with chronic low back pain: protocol of a retrospective longitudinal study / Allegri M., Gregori M. De, Minella C.E.,

Klersy C., Wang W., Sim M., Gieger C., Manz J., Pemberton I.K., MacDougall J., Williams F.M., Zundert J. Van, Buyse K., Lauc G., Gudelj I., Primorac D., Skelin A., Aulchenko Y.S., Karssen L.C., Kapural L., Rauck R., Fanelli G. // BMJ Open – 2016. – T. 6 – № 10 – C.e012070.

113. Theodoratou E. Dietary Vitamin B6 Intake and the Risk of Colorectal Cancer / Theodoratou E., Farrington S.M., Tenesa A., McNeill G., Cetnarskyj R., Barnetson R.A., Porteous M.E., Dunlop M.G., Campbell H. // Cancer Epidemiology Biomarkers & Prevention – 2008. – T. 17 – № 1 – C.171–182.

114. COGENT Study Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer / COGENT Study, Houlston R.S., Webb E., Dunlop M.G. // Nature Genetics – 2008. – T. 40 – № 12 – C.1426–1435.

115. Dunlop M.G. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk / Dunlop M.G., Dobbins S.E., Farrington S.M., Houlston R.S. // Nature Genetics – 2012. – T. 44 – № 7 – C.770–776.

116. Tillin T. Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. / Tillin T., Forouhi N.G., McKeigue P.M., Chaturvedi N., SABRE Study Group // International journal of epidemiology – 2012. – T. 41 – № 1 – C.33–42.

117. Loh P.R. Fast and accurate long-range phasing in a UK Biobank cohort / Loh P.R., Palamara P.F., Price A.L. // Nature Genetics – 2016. – T. 48 – № 7 – C.811–816.

118. Das S. Next-generation genotype imputation service and methods / Das S., Forer L., Schönherr S., Sidore C., Locke A.E., Kwong A., Vrieze S.I., Chew E.Y., Levy S., McGue M., Schlessinger D., Stambolian D., Loh P.R., Iacono W.G., Swaroop A., Scott L.J., Cucca F., Kronenberg F., Boehnke M., Abecasis G.R., Fuchsberger C. // Nature Genetics – 2016. – T. 48 – № 10 – C.1284–1287.

119. Vuckovic F. IgG Glycome in Colorectal Cancer / Vuckovic F., Theodoratou

E., Thaci K., Timofeeva M., Vojta A., Stambuk J., Pucic-Bakovic M., Rudd P.M., Erek L., Servis D., Wennerstro m A., Farrington S.M., Perola M., Aulchenko Y., Dunlop M.G., Campbell H., Lauc G. // Clinical Cancer Research – 2016. – Т. 22 – № 12 – С.3078–3086.

120. Theodoratou E. Glycosylation of plasma IgG in colorectal cancer prognosis / Theodoratou E., Thaçi K., Agakov F., Timofeeva M.N., Štambuk J., Pučić-Baković M., Vučković F., Orchard P., Agakova A., Din F.V.N., Brown E., Rudd P.M., Farrington S.M., Dunlop M.G., Campbell H., Lauc G. // Scientific Reports – 2016. – Т. 6 – № 1 – С.28098.

121. Delaneau O. Improved whole-chromosome phasing for disease and population genetic studies / Delaneau O., Zagury J.-F., Marchini J. // Nature Methods – 2013. – Т. 10 – № 1 – С.5–6.

122. Suhre K. Fine-Mapping of the Human Blood Plasma N-Glycome onto Its Proteome / Suhre K., Trbojević-Akmačić I., Ugrina I., Mook-Kanamori D., Spector T., Graumann J., Lauc G., Falchi M. // Metabolites – 2019. – Т. 9 – № 7 – С.122.

123. Johnson W.E. Adjusting batch effects in microarray expression data using empirical Bayes methods / Johnson W.E., Li C., Rabinovic A. // Biostatistics – 2007. – Т. 8 – № 1 – С.118–127.

124. Sharapov S.Z. Defining the genetic control of human blood plasma N-glycome using genome-wide association study / Sharapov S.Z., Tsepilov Y.A., Klaric L., Mangino M., Thareja G., Shadrina A.S., Simurina M., Dagostino C., Dmitrieva J., Vilaj M., Vuckovic F., Pavic T., Stambuk J., Trbojevic-Akmacic I., Kristic J., Simunovic J., Momcilovic A., Campbell H., Doherty M., Dunlop M.G., Farrington S.M., Pucic-Bakovic M., Gieger C., Allegri M., Louis E., Georges M., Suhre K., Spector T., Williams F.M.K.K., Lauc G., Aulchenko Y.S. // Human Molecular Genetics – 2019. – Т. 28 – № 12 – С.2062–2077.

125. Benedetti E. Systematic Evaluation of Normalization Methods for Glycomics Data Based on Performance of Network Inference / Benedetti E., Gerstner N., Pučić-

Baković M., Keser T., Reiding K.R., Ruhaak L.R., Štambuk T., Selman M.H.J., Rudan I., Polašek O., Hayward C., Beekman M., Slagboom E., Wuhrer M., Dunlop M.G., Lauc G., Krumsiek J. // Metabolites – 2020. – Т. 10 – № 7 – С.271.

126. Feingold E. Regression-Based Quantitative-Trait–Locus Mapping in the 21st Century / Feingold E. // The American Journal of Human Genetics – 2002. – Т. 71 – № 2 – С.217–222.

127. Zhou X. Efficient multivariate linear mixed model algorithms for genome-wide association studies / Zhou X., Stephens M. // Nature Methods – 2014. – Т. 11 – № 4 – С.407–409.

128. Devlin B. Genomic control for association studies. / Devlin B., Roeder K. // Biometrics – 1999. – Т. 55 – № 4 – С.997–1004.

129. Willer C.J. METAL: fast and efficient meta-analysis of genomewide association scans / Willer C.J., Li Y., Abecasis G.R. // Bioinformatics – 2010. – Т. 26 – № 17 – С.2190–2191.

130. Schaid D.J. From genome-wide associations to candidate causal variants by statistical fine-mapping // Nat. Rev. Genet. – 2018. – Т. 19. – № 8.

131. Bunt M. van de Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci / Bunt M. van de, Cortes A., Brown M.A., Morris A.P., McCarthy M.I. // PLOS Genetics – 2015. – Т. 11 – № 9 – С.e1005535.

132. Stenson P.D. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies / Stenson P.D., Mort M., Ball E. V, Evans K., Hayden M., Heywood S., Hussain M., Phillips A.D., Cooper D.N. // Human Genetics – 2017. – Т. 136 – № 6 – С.665–677.

133. International HapMap 3 Consortium, Altshuler D.M. Integrating common and rare genetic variation in diverse human populations / International HapMap 3 Consortium, Altshuler D.M., Gibbs R.A., Peltonen L., McEwen J.E. // Nature – 2010. –

T. 467 – № 7311 – C.52–58.

134. Gibbs R.A. A global reference for human genetic variation / Gibbs R.A., Boerwinkle E., Doddapaneni H., Rasheed A. // Nature – 2015. – T. 526 – № 7571 – C.68–74.

135. Bush W.S. Chapter 11: Genome-Wide Association Studies / Bush W.S., Moore J.H. // PLoS Computational Biology – 2012. – T. 8 – № 12 – C.e1002822.

136. Sharapov S.Z. Replication of 15 loci involved in human plasma protein N-glycosylation in 4802 samples from four cohorts / Sharapov S.Z., Shadrina A.S., Tsepilov Y.A., Elgaeva E.E., Tiys E.S., Feoktistova S.G., Zaytseva O.O., Vuckovic F., Cuadrat R., Jäger S., Wittenbecher C., Karssen L.C., Timofeeva M., Tillin T., Trbojević-Akmačić I., Štambuk T., Rudman N., Krištić J., Šimunović J., Momčilović A., Vilaj M., Jurić J., Slana A., Gudelj I., Klarić T., Puljak L., Skelin A., Kadić A.J., Zundert J. Van, Chaturvedi N., Campbell H., Dunlop M., Farrington S.M., Doherty M., Dagostino C., Gieger C., Allegri M., Williams F., Schulze M.B., Lauc G., Aulchenko Y.S. // Glycobiology – 2021. – T. 31 – № 2 – C.82–88.

137. Adzhubei I. Predicting Functional Effect of Human Missense Mutations Using PolyPhen- 2 / Adzhubei I., Jordan D.M., Sunyaev S.R. // Current Protocols in Human Genetics – 2013. – T. 76 – № 1.

138. Kumar P. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm / Kumar P., Henikoff S., Ng P.C. // Nature Protocols – 2009. – T. 4 – № 7 – C.1073–1081.

139. Mollicone R. Molecular basis for plasma alpha(1,3)-fucosyltransferase gene deficiency (FUT6). / Mollicone R., Reguigne I., Fletcher A., Aziz A., Rustam M., Weston B.W., Kelly R.J., Lowe J.B., Oriol R. // The Journal of biological chemistry – 1994. – T. 269 – № 17 – C.12662–12671.

140. Puan K.J. FUT6 deficiency compromises basophil function by selectively abrogating their sialyl-Lewis x expression / Puan K.J., San Luis B., Yusof N., Rötzschke

O. // Communications Biology – 2021. – T. 4 – № 1 – C.832.

141. Astle W.J. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease / Astle W.J., Elding H., Jiang T., Soranzo N. // Cell – 2016. – T. 167 – № 5 – C.1415–1429.

142. Fritsche L.G. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants / Fritsche L.G., Igl W., Bailey J.N.C., Heid I.M. // Nature Genetics – 2016. – T. 48 – № 2 – C.134–143.

143. Fritsche L.G. Seven new loci associated with age-related macular degeneration. / Fritsche L.G., Chen W., Schu M., AMD Gene Consortium // Nature genetics – 2013. – T. 45 – № 4 – C.433–9, 439e1-2.

144. Willer C.J. Discovery and refinement of loci associated with lipid levels / Willer C.J., Schmidt E.M., Sengupta S., Abecasis G.R. // Nature Genetics – 2013. – T. 45 – № 11 – C.1274–1283.

145. Teslovich T.M. Biological, clinical and population relevance of 95 loci for blood lipids / Teslovich T.M., Musunuru K., Smith A. V., Kathiresan S. // Nature – 2010. – T. 466 – № 7307 – C.707–713.

146. Prins B.P. Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations. / Prins B.P., Kuchenbaecker K.B., Bao Y., Smart M., Zabaneh D., Fatemifar G., Luan J., Wareham N.J., Scott R.A., Perry J.R.B., Langenberg C., Benzeval M., Kumari M., Zeggini E. // Scientific reports – 2017. – T. 7 – № 1 – C.11008.

147. Ligthart S. Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation. / Ligthart S., Vaez A., Hsu Y.-H., Inflammation Working Group of the CHARGE Consortium, PMI-WG-XCP, LifeLines Cohort Study, Stolk R., Uitterlinden A.G., Hofman A., Alizadeh B.Z., Franco O.H., Dehghan A. // BMC genomics – 2016. – T. 17 – C.443.

148. Scott R.A. An Expanded Genome-Wide Association Study of Type 2 Diabetes

in Europeans. / Scott R.A., Scott L.J., Mägi R., DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium // Diabetes – 2017. – Т. 66 – № 11 – С.2888–2902.

149. Chambers J.C. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma / Chambers J.C., Zhang W., Sehmi J., Kooner J.S. // Nature Genetics – 2011. – Т. 43 – № 11 – С.1131–1138.

150. Perry J.R.B. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche / Perry J.R.B., Day F., Elks C.E., Ong K.K. // Nature – 2014. – Т. 514 – № 7520 – С.92–97.

151. Wood A.R. Defining the role of common variation in the genomic and biological architecture of adult human height / Wood A.R., Esko T., Yang J., Frayling T.M. // Nature Genetics – 2014. – Т. 46 – № 11 – С.1173–1186.

152. Gregersen P.K. REL, encoding a member of the NF-κB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis / Gregersen P.K., Amos C.I., Lee A.T., Lu Y., Remmers E.F., Kastner D.L., Seldin M.F., Criswell L.A., Plenge R.M., Holers V.M., Mikuls T.R., Sokka T., Moreland L.W., Bridges S.L., Xie G., Begovich A.B., Siminovitch K.A. // Nature Genetics – 2009. – Т. 41 – № 7 – С.820–823.

153. Klarić L. Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases / Klarić L., Tsepilov Y.A., Stanton C.M., Hayward C. // Science Advances – 2020. – Т. 6 – № 8 – С.eaax0301.

154. Brockhausen I. Glycosyltransferases Involved in N–and O–Glycan Biosynthesis / Brockhausen I., Schachter H. // Glycosciences – 1996. – С.79–113.

155. Dupuy F. α1,4-Fucosyltransferase Activity: A Significant Function in the Primate Lineage has Appeared Twice Independently / Dupuy F., Germot A., Marenda M., Oriol R., Blancher A., Julien R., Maftah A. // Molecular Biology and Evolution – 2002. – Т. 19 – № 6 – С.815–824.

156. Taniguchi N.Handbook of glycosyltransferases and related genes, second edition / N. Taniguchi, K. Honke, M. Fukuda, H. Narimatsu, Y. Yamaguchi, T. Angata –

, 2014.

157. Staretz-Chacham O. B4GALT1-congenital disorders of glycosylation: Expansion of the phenotypic and molecular spectrum and review of the literature. / Staretz-Chacham O., Noyman I., Wormser O., Abu Quider A., Hazan G., Morag I., Hadar N., Raymond K., Birk O.S., Ferreira C.R., Koifman A. // Clinical genetics – 2020. – Т. 97 – № 6 – С.920–926.

158. Pottier N. Expression of SMARCB1 modulates steroid sensitivity in human lymphoblastoid cells: identification of a promoter snp that alters PARP1 binding and SMARCB1 expression / Pottier N., Cheok M.H., Yang W., Assem M., Tracey L., Obenauer J.C., Panetta J.C., Relling M. V, Evans W.E. // Human Molecular Genetics – 2007. – Т. 16 – № 19 – С.2261–2271.

159. Oda Y. Derlin-2 and Derlin-3 are regulated by the mammalian unfolded protein response and are required for ER-associated degradation / Oda Y., Okada T., Yoshida H., Kaufman R.J., Nagata K., Mori K. // The Journal of Cell Biology – 2006. – Т. 172 – № 3 – С.383–393.

160. Martinvalet D. The role of the mitochondria and the endoplasmic reticulum contact sites in the development of the immune responses / Martinvalet D. // Cell Death & Disease – 2018. – Т. 9 – № 3 – С.336.

161. Kellokumpu S. Golgi pH, ion and redox homeostasis: How much do they really matter? / Kellokumpu S. // Frontiers in Cell and Developmental Biology – 2019. – Т. 7.

162. Hassinen A. Functional organization of Golgi N- and O-glycosylation pathways involves pH-dependent complex formation that is impaired in cancer cells / Hassinen A., Pujol F.M., Kokkonen N., Pieters C., Kihlström M., Korhonen K., Kellokumpu S. // Journal of Biological Chemistry – 2011. – Т. 286 – № 44 – С.38329–38340.

163. Marke R. The many faces of IKZF1 in B-cell precursor acute lymphoblastic

leukemia / Marke R., Leeuwen F.N. van, Scheijen B. // Haematologica – 2018. – T. 103 – № 4 – C.565–574.

164. Sellars M. Ikaros controls isotype selection during immunoglobulin class switch recombination / Sellars M., Reina-San-Martin B., Kastner P., Chan S. // Journal of Experimental Medicine – 2009. – T. 206 – № 5 – C.1073–1087.

165. Kennedy A.E. What has GWAS done for HLA and disease associations? / Kennedy A.E., Ozbek U., Dorak M.T. // International Journal of Immunogenetics – 2017. – T. 44 – № 5 – C.195–211.

166. Shen X. Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation / Shen X., Klarić L., Sharapov S., Mangino M., Ning Z., Wu D., Trbojević-Akmačić I., Pučić-Baković M., Rudan I., Polašek O., Hayward C., Spector T.D., Wilson J.F., Lauc G., Aulchenko Y.S. // Nature Communications – 2017. – T. 8 – № 1 – C.447.

167. Ahn J. Separation of 2-aminobenzamide labeled glycans using hydrophilic interaction chromatography columns packed with 1.7 μm sorbent / Ahn J., Bones J., Yu Y.Q., Rudd P.M., Gilar M. // Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences – 2010. – T. 878 – № 3–4 – C.403–408.

Приложение

Доп. табл. 1. **Сравнение результатов гармонизации пиков методами суммирования площадей и интеграции на хроматограмме.** Гармонизация пиков GP24 и GP25 в пик PGP22 была выполнена двумя методами для 35 образцов плазмы из 3 выборок: PainOR, TwinsUK, QMDIab. PGP22_сумм - значение гармонизированного пика PGP22, полученного суммированием площадей пиков GP24 и GP25. PGP22_интег - значение гармонизированного пика PGP22, полученного интегрированием пиков GP24 и GP25 на хроматограмме.

Номер образца	PGP22_сум	PGP22_интег
1	1243382	1243383
2	882107	882107
3	2155160	2155160
4	889547	889547
5	830644	830644
6	818061	818062
7	961162	961161
8	1498924	1498924
9	887204	887204
10	1532743	1532743
11	1423928	1423928
12	1753836	1753836
13	1543883	1543884
14	1428139	1428139
15	328564	328564
16	1121500	1121500
17	780756	780756
18	654441	654441
19	1355445	1355445
20	1041631	1041631
21	409267	409267
22	1100345	1100345
23	664148	664147
24	543274	543274
25	805835	805836
26	1014660	1014660
27	655317	655318
28	1528045	1528045
29	1347073	1347074
30	375288	375288
31	742619	742618
32	974831	974830
33	1014524	1014524
34	248207	248207
35	295361	295360

Доп. табл. 2. Описание 36 гликомных признаков и 77 производных признаков. Название признака – краткое название признака. Описание – биохимическое описание признака. Формула расчета – формула расчета признака на основе 36 гармонизированных признаков.

Название признака	Описание	Формула расчета
PGP1	The percentage of FA2 in total plasma glycans	GP1 / GP * 100
PGP2	The percentage of M5 + FA2B in total plasma glycans	GP2 / GP * 100
PGP3	The percentage of A2[6]BG1 in total plasma glycans	GP3 / GP * 100
PGP4	The percentage of FA2[6]G1 in total plasma glycans	GP4 / GP * 100
PGP5	The percentage of FA2[3]G1 in total plasma glycans	GP5 / GP * 100
PGP6	The percentage of FA2[6]BG1 in total plasma glycans	GP6 / GP * 100
PGP7	The percentage of M6 D3 in total plasma glycans	GP7 / GP * 100
PGP8	The percentage of A2G2 in total plasma glycans	GP8 / GP * 100
PGP9	The percentage of A2BG2 in total plasma glycans	GP9 / GP * 100
PGP10	The percentage of FA2G2 in total plasma glycans	GP10 / GP * 100
PGP11	The percentage of FA2BG2 in total plasma glycans	GP11 / GP * 100
PGP12	The percentage of A2[3]BG1S[3]1 + A2[3]BG1S[6]1 in total plasma glycans	GP12 / GP * 100
PGP13	The percentage of FA2[3]G1S[3]1 + FA2[3]G1S[6]1 in total plasma glycans	GP13 / GP * 100
PGP14	The percentage of A2G2S[6]1 + A2G2S[3]1 in total plasma glycans	GP14 / GP * 100
PGP15	The percentage of FA2G2S[6]1 + FA2G2S[3]1 in total plasma glycans	GP15 / GP * 100
PGP16	The percentage of FA2BG2S[3]1 + FA2BG2S[3]1 in total plasma glycans	GP16 / GP * 100
PGP17	The percentage of A2G2S[3,6]2 + A2G2S[3,3]2 in total plasma glycans	GP17 / GP * 100
PGP18	The percentage of M9 in total plasma glycans	GP18 / GP * 100
PGP19	The percentage of A2G2S[3,6]2 + A2G2S[6,6]2 + A2G2S[3,3]2 in total plasma glycans	GP19 / GP * 100
PGP20	The percentage of FA2G2S[3,6]2 + FA2G2S[3,3]2 + FA2G2S[6,6]2 in total plasma glycans	GP20 / GP * 100
PGP21	The percentage of FA2BG2S[3,6]2 + FA2BG2S[6,6]2 + FA2BG2S[3,3]2 in total plasma glycans	GP21 / GP * 100
PGP22	The percentage of A3G3S[3,6]2 in total plasma glycans	GP22 / GP * 100
PGP23	The percentage of A3G3S[3,3]2 in total plasma glycans	GP23 / GP * 100
PGP24	The percentage of A3G3S[3,3,3]3 in total plasma glycans	GP24 / GP * 100
PGP25	The percentage of A3G3S[3,3,6]3 in total plasma glycans	GP25 / GP * 100
PGP26	The percentage of FA3G3S[3,3,3]3 in total plasma glycans	GP26 / GP * 100
PGP27	The percentage of A3G3S[3,3,6]3 + A3G3S[3,6,6]3 in total plasma glycans	GP27 / GP * 100

PGP28	The percentage of FA3G3S[3,3,6]3 + FA3G3S[3,6,6]3 in total plasma glycans	GP28 / GP * 100
PGP29	The percentage of A3F1G3S[3,3,3]3 + A3F1G3S[3,3,6]3in total plasma glycans	GP29 / GP * 100
PGP30	The percentage of A4G4S[3,3,3]3 in total plasma glycans	GP30 / GP * 100
PGP31	The percentage of A4G4S[3,3,6]3 + A4G4S[3,6,6]3 in total plasma glycans	GP31 / GP * 100
PGP32	The percentage of A4F1G3S[3,3,3]3 + A4F1G3S[3,3,6]3 + A4F1G3S[3,6,6]3 in total plasma glycans	GP32 / GP * 100
PGP33	The percentage of A4G4S[3,3,3,3]4 in total plasma glycans	GP33 / GP * 100
PGP34	The percentage of A4G4S[3,3,3,6]4 in total plasma glycans	GP34 / GP * 100
PGP35	The percentage of A4G4S[3,6,6,6]4 in total plasma glycans	GP35 / GP * 100
PGP36	The percentage of A4F1G4S[3,3,3,6]4 in total plasma glycans	GP36 / GP * 100
PGP37	The percentage of sialylation of core-fucosylated galactosylated structures without bisecting GlcNAc in total plasma glycans	SUM(GP13+GP15+GP20+GP26+GP28) / SUM(GP4+GP5+GP10+GP13+GP15+GP20+GP26+GP28) * 100
PGP38	The percentage of sialylation of core-fucosylated galactosylated structures with bisecting GlcNAc in total plasma glycans	SUM(GP16+GP21) / SUM(GP6+GP11+GP16+GP21) * 100
PGP39	The percentage of sialylation of all core-fucosylated structures without bisecting GlcNAc in total plasma glycans	SUM(GP13+GP15+GP20+GP26+GP28) / SUM(GP1+GP4+GP5+GP10+GP13+GP15+GP20+GP26+GP28) * 100
PGP40	The percentage of sialylation of all core-fucosylated structures with bisecting GlcNAc in total plasma glycans	SUM(GP16+GP21) / SUM(GP2+GP6+GP11+GP16+GP21) * 100
PGP41	The percentage of monosialylation of core-fucosylated monogalactosylated structures without bisecting GlcNAc in total plasma glycans	GP13 / SUM(GP4+GP5+GP13) * 100
PGP42	The percentage of monosialylation of core-fucosylated digalactosylated structures without bisecting GlcNAc in total plasma glycans	GP15 / SUM(GP10+GP15 + GP20) * 100
PGP43	The percentage of disialylation of core-fucosylated digalactosylated structures without bisecting GlcNAc in total plasma glycans	GP20 / SUM (GP10+GP15+GP20) * 100
PGP44	The percentage of monosialylation of core-fucosylated digalactosylated structures with bisecting GlcNAc in total plasma glycans	GP16 / SUM (GP11+GP16+GP21) * 100
PGP45	The percentage of disialylation of core-fucosylated digalactosylated structures with bisecting GlcNAc in total plasma glycans	GP21 / SUM (GP11+GP16+GP21) * 100

PGP46	Ratio of all fucosylated monosialylated and disialylated structures (+/- bisecting GlcNAc) in total plasma glycans	SUM(GP13+GP15+GP16) / SUM(GP20+GP21)
PGP47	Ratio of fucosylated monosialylated and disialylated structures (without bisecting GlcNAc) in total plasma glycans	SUM(GP13+GP15) / GP20
PGP48	Ratio of fucosylated monosialylated and disialylated structures (with bisecting GlcNAc) in total plasma glycans	GP16 / GP21
PGP49	Ratio of all core-fucosylated monosialylated and trisialylated structures (+/- bisecting GlcNAc) in total plasma glycans	SUM(GP13+GP15+GP16) / SUM(GP26+GP28)
PGP50	Ratio of core-fucosylated monosialylated and trisialylated structures (without bisecting GlcNAc) in total plasma glycans	SUM(GP13+GP15) / SUM(GP26+GP28)
PGP51	Ratio of all core-fucosylated disialylated and trisialylated structures (+/- bisecting GlcNAc) in total plasma glycans	SUM(GP20+GP21) / SUM(GP26+GP28)
PGP52	Ratio of core-fucosylated disialylated and trisialylated structures (without bisecting GlcNAc) in total plasma glycans	SUM(GP20) / SUM(GP26+GP28)
PGP53	Ratio of all core-fucosylated sialylated structures with and without bisecting GlcNAc in total plasma glycans	SUM(GP16+GP21) / SUM(GP13+GP15+GP20+GP26+GP28)
PGP54	Ratio of fucosylated monosialylated structures with and without bisecting GlcNAc in total plasma glycans	GP16 / SUM(GP13+GP15)
PGP55	The incidence of bisecting GlcNAc in all fucosylated monosialylated structures in total plasma glycans	GP16 / SUM(GP13+GP15+GP16)
PGP56	Ratio of fucosylated disialylated structures with and without bisecting GlcNAc in total plasma glycans	GP21 / GP20
PGP57	The incidence of bisecting GlcNAc in all fucosylated disialylated structures in total plasma glycans	GP21 / SUM(GP20+GP21)
PGP58	The percentage of FA2 in total neutral plasma glycans (GPn)	GP1 / GPn * 100
PGP59	The percentage of M5 + FA2B in total neutral plasma glycans (GPn)	GP2 / GPn * 100
PGP60	The percentage of A2[6]BG1 in total neutral plasma glycans (GPn)	GP3 / GPn * 100
PGP61	The percentage of FA2[6]G1 in total neutral plasma glycans (GPn)	GP4 / GPn * 100
PGP62	The percentage of FA2[3]G1 in total neutral plasma glycans (GPn)	GP5 / GPn * 100
PGP63	The percentage of FA2[6]BG1 in total neutral plasma glycans (GPn)	GP6 / GPn * 100
PGP64	The percentage of M6 D3 in total neutral plasma glycans (GPn)	GP7 / GPn * 100
PGP65	The percentage of A2G2 in total neutral plasma glycans (GPn)	GP8 / GPn * 100
PGP66	The percentage of A2BG2 in total neutral plasma glycans (GPn)	GP9 / GPn * 100

PGP67	The percentage of FA2G2 in total neutral plasma glycans (GPn)	GP10 / GPn * 100
PGP68	The percentage of FA2BG2 in total neutral plasma glycans (GPn)	GP11 / GPn * 100
PGP69	The percentage of M9 in total neutral plasma glycans (GPn)	GP18 / GPn * 100
PGP70	The percentage of agalactosylated structures in total neutral plasma glycans	GP1n+GP2n
PGP71	The percentage of monogalactosylated structures in total neutral plasma glycans	GP3n+GP4n+GP5n+GP6n
PGP72	The percentage of digalactosylated structures in total neutral plasma glycans	GP8n+GP9n+GP10n+GP11n
PGP73	The percentage of all fucosylated structures (+/- bisecting GlcNAc) in total neutral plasma glycans	GP1n+GP2n+GP4n+GP5n+GP6n+GP10n+GP11n
PGP74	The percentage of fucosylation of monogalactosylated structures in total neutral plasma glycans	(GP4n+GP5n+GP6n) / G1n * 100
PGP75	The percentage of fucosylation of digalactosylated structures in total neutral plasma glycans	(GP10n+GP11n) / G2n * 100
PGP76	The percentage of fucosylated structures (without bisecting GlcNAc) in total neutral plasma glycans	GP1n+GP4n+GP5n+GP10n
PGP77	The percentage of fucosylation of agalactosylated structures (without bisecting GlcNAc) in total neutral plasma glycans	GP1n / G0n * 100
PGP78	The percentage of fucosylation of monogalactosylated structures (without bisecting GlcNAc) in total neutral plasma glycans	(GP4n+GP5n) / G1n * 100
PGP79	The percentage of fucosylation of digalactosylated structures (without bisecting GlcNAc) in total neutral plasma glycans	GP10n / G2n * 100
PGP80	The percentage of fucosylated structures (with bisecting GlcNAc) in total neutral plasma glycans	GP2n+GP6n+GP11n
PGP81	The percentage of fucosylation of agalactosylated structures (with bisecting GlcNAc) in total neutral plasma glycans	GP2n / G0n * 100
PGP82	The percentage of fucosylation of monogalactosylated structures (with bisecting GlcNAc) in total neutral plasma glycans	GP6n / G1n * 100
PGP83	The percentage of fucosylation of digalactosylated structures (with bisecting GlcNAc) in total neutral plasma glycans	GP11n / G2n * 100
PGP84	Ratio of fucosylated structures with and without bisecting GlcNAc in total neutral plasma glycans	FBn / Fn
PGP85	The incidence of bisecting GlcNAc in all fucosylated structures in total neutral plasma glycans	FBn / Fn total * 100
PGP86	Ratio of fucosylated non-bisecting GlcNAc structures and all structures with bisecting GlcNAc in total neutral plasma glycans	Fn / (FBn + GP3n + GP9n)
PGP87	Ratio of afucosylated structures with bisecting GlcNAc and all fucosylated structures (+/- bisecting GlcNAc) in total neutral plasma glycans	(GP3n + GP9n) / (Fn + Fbn) * 100

PGP88	Ratio of fucosylated digalactosylated structures with and without bisecting GlcNAc in total neutral plasma glycans	GP11n / GP10n
PGP89	The incidence of bisecting GlcNAc in all fucosylated digalactosylated structures in total neutral plasma glycans	GP11n / (GP10n + GP11n) * 100
PGP90	Ratio of fucosylated digalactosylated non-bisecting GlcNAc structures and all digalactosylated structures with bisecting GlcNAc in total neutral plasma glycans	GP10n / (GP9n + GP11n)
PGP91	Ratio of afucosylated digalactosylated structures with bisecting GlcNAc and all fucosylated digalactosylated structures (+/- bisecting GlcNAc) in total neutral plasma glycans	GP9n / (GP10n + GP11n)
PGP92	The percentage of antennary fucosylated structures in total plasma glycome	GP29+GP32+GP36
PGP93	The percentage of core fucosylated structures in total plasma glycome	GP1+GP2+GP4+GP5+GP6+GP10+GP11+GP13+GP15+GP16+GP20+GP21+GP26+GP28
PGP94	The percentage of neutral glycan structures in total plasma glycome	GP1+GP2+GP3+GP4+GP5+GP6+GP7+GP8+GP9+GP10+GP11
PGP95	the percentage of monosyalated structures in total plasma glycome	GP12+GP13+GP14+GP15+GP16
PGP96	the percentage of bisyalated structures in total plasma glycome	GP17+GP19+GP20+GP21+GP22+GP23
PGP97	the percentage of trisyalated structures in total plasma glycome	GP24+GP25+GP26+GP27+GP28+GP29+GP30+GP31+GP32
PGP98	the percentage of tetrasyalated structures in total plasma glycome	GP33+GP34+GP35+GP36
PGP99	The percentage of agalactosylated structures in total plasma glycans	GP1+GP2
PGP100	The percentage of monogalactosylated structures in total plasma glycans	GP3+GP4+GP5+GP6+GP12+GP13
PGP101	The percentage of digalactosylated structures in total plasma glycans	GP8+GP9+GP10+GP11+GP14+GP15+GP16+GP17+GP19+GP20+GP21
PGP102	The percentage of trigalactosylated structures in total plasma glycans	GP22+GP23+GP24+GP25+GP26+GP27+GP28+GP29+GP32
PGP103	The percentage of tetragalactosylated structures in total plasma glycans	GP30+GP31+GP33+GP34+GP35+GP36
PGP104	The percentage of biantennary structures in total plasma glycans	GP1+GP2+GP3+GP4+GP5+GP6+GP8+GP9+GP10+GP11+GP12+GP13+GP14+GP15+GP16+GP17+GP19+GP20+GP21
PGP105	The percentage of triiantennary structures in total plasma glycans	GP22+GP23+GP24+GP25+GP26+GP27+GP28+GP29
PGP106	The percentage of tetraantenary structures in total plasma glycans	GP30+GP31+GP32+GP33+GP34+GP35+GP36
PGP107	The percentage of high-mannose structures in total plasma glycans	GP2+GP7+GP18

PGP108	The percentage of glycan structures with bisecting GlcNAc in total plasma glycans	GP2+GP3+GP6+GP9+GP11+ GP12+GP16+GP21
PGP109	Ratio of disialylated and trisialylated trigalactosylated structures in total plasma glycans	SUM(GP22+GP23) / SUM(GP24+GP25+GP26+GP27+GP28+GP29+GP32)
PGP110	Ratio of trisialylated and tetrasialylated tetragalactosylated structures in total plasma glycans	SUM(GP30+GP31) / SUM(GP33+GP34+GP35+GP36)
PGP111	The percentage of core-fucosylation of trigalactosylated structures in total plasma glycans	SUM(GP26+GP28) / SUM(GP22+GP23+GP24+GP25+GP26+GP27+GP28+GP29+ GP32) * 100
PGP112	The percentage of antennary-fucosylation of trigalactosylated structures in total plasma glycans	SUM(GP29+GP32) / SUM(GP22+GP23+GP24+GP25+GP26+GP27+GP28+GP29+ GP32) * 100
PGP113	The percentage of antennary-fucosylation of tetragalactosylated structures in total plasma glycans	GP36 / SUM(GP30+GP31+GP33+GP34+GP35+GP36) * 100

Доп. табл. 3. Описание 36 гликомных признаков и 81 производного признака. Название признака – краткое название признака. Описание – биохимическое описание признака. Формула расчета – формула расчета признака на основе 36 гармонизированных признаков.

Название признака	Описание	Формула расчета
PGP1	The percentage of FA2	GP1
PGP2	The percentage of FA2B	GP2
PGP3	The percentage of A2BG1	GP3
PGP4	The percentage of FA2G1	GP4
PGP5	The percentage of FA2G1	GP5
PGP6	The percentage of FA2BG1	GP6
PGP7	The percentage of M6	GP7
PGP8	The percentage of A2G2	GP8
PGP9	The percentage of A2BG2	GP9
PGP10	The percentage of FA2G2	GP10
PGP11	The percentage of FA2BG2	GP11
PGP12	The percentage of M7+A2G2S1	GP12
PGP13	The percentage of FA2G1S1	GP13
PGP14	The percentage of A2G2S1+A2G2S1	GP14 + GP15
PGP15	The percentage of FA2G2S1+FA2G2S1	GP16
PGP16	The percentage of FA2BG2S1+FA2BG2S1	GP17
PGP17	The percentage of A2G2S2	GP18
PGP18	The percentage of M9	GP19
PGP19	The percentage of A2G2S2	GP20+GP21
PGP20	The percentage of FA2G2S2	GP22
PGP21	The percentage of FA2BG2S2	GP23
PGP22	The percentage of A3G3S2	GP24+GP25
PGP23	The percentage of A3G3S2	GP26
PGP24	The percentage of A3F1G3S2	GP27
PGP25	The percentage of A3G3S3	GP28
PGP26	The percentage of A3G3S3	GP29
PGP27	The percentage of A3G3S3	GP30
PGP28	The percentage of FA3G3S3	GP31
PGP29	The percentage of A3G3S3	GP32
PGP30	The percentage of A3F1G3S3	GP33
PGP31	The percentage of FA3G3S3	GP34
PGP32	The percentage of FA3F1G3S3	GP35
PGP33	The percentage of A4G4S3	GP36
PGP34	The percentage of A4G4S4	GP37
PGP35	The percentage of A4G4S4	GP38
PGP36	The percentage of A4F1G4S4	GP39
PGP37	The percentage of sialylation of core-fucosylated galactosylated structures without bisecting GlcNAc	SUM(PGP13+PGP15+PGP20+PGP28+PGP31+PGP32)/SUM(PGP4+PGP5+PGP10+PGP13+PGP15+PGP20+PGP28+PGP31+PGP32)*100
PGP38	The percentage of sialylation of core-fucosylated galactosylated structures with bisecting GlcNAc	SUM(PGP16+PGP21)/SUM(PGP6+PGP11+PGP16+PGP21)*100

PGP39	The percentage of sialylation of all core-fucosylated structures without bisecting GlcNAc	SUM(PGP13+PGP15+PGP20+PGP28+PGP31+PGP32)/SUM(PGP1+PGP4+PGP5+PGP10+PGP13+PGP15+PGP20+PGP28+PGP31+PGP32)*100
PGP40	The percentage of sialylation of all core-fucosylated structures with bisecting GlcNAc	SUM(PGP16+PGP21)/SUM(PGP2+PGP6+PGP11+PGP16+PGP21)*100
PGP41	The percentage of monosialylation of core-fucosylated monogalactosylated structures without bisecting GlcNAc	PGP13/SUM(PGP4+PGP5+PGP13)*100
PGP42	The percentage of monosialylation of core-fucosylated digalactosylated structures without bisecting GlcNAc	PGP15/SUM(PGP10+PGP15+PGP20)*100
PGP43	The percentage of disialylation of core-fucosylated digalactosylated structures without bisecting GlcNAc	PGP20/SUM(PGP10+PGP15+PGP20)*100
PGP44	The percentage of monosialylation of core-fucosylated digalactosylated structures with bisecting GlcNAc	PGP16/SUM(PGP11+PGP16+PGP21)*100
PGP45	The percentage of disialylation of core-fucosylated digalactosylated structures with bisecting GlcNAc	PGP21/SUM(PGP11+PGP16+PGP21)*100
PGP46	Ratio of all fucosylated monosialylated and disialylated structures (+/- bisecting GlcNAc)	SUM(PGP13+PGP15+PGP16)/SUM(PGP20+PGP21+PGP24)
PGP47	Ratio of fucosylated monosialylated and disialylated structures (without bisecting GlcNAc)	SUM(PGP13+PGP15)/SUM(PGP20+PGP24)
PGP48	Ratio of fucosylated monosialylated and disialylated structures (with bisecting GlcNAc)	PGP16/PGP21
PGP49	Ratio of all core-fucosylated monosialylated and trisialylated structures (+/- bisecting GlcNAc)	SUM(PGP13+PGP15+PGP16)/SUM(PGP28+PGP31+PGP32)
PGP50	Ratio of core-fucosylated monosialylated and trisialylated structures (without bisecting GlcNAc)	SUM(PGP13+PGP15)/SUM(PGP28+PGP31+PGP32)
PGP51	Ratio of all core-fucosylated disialylated and trisialylated structures (+/- bisecting GlcNAc)	SUM(PGP20+PGP21)/SUM(PGP28+PGP31+PGP32)
PGP52	Ratio of core-fucosylated disialylated and trisialylated structures (without bisecting GlcNAc)	PGP20/SUM(PGP28+PGP31+PGP32)
PGP53	Ratio of all core-fucosylated sialylated structures with and without bisecting GlcNAc	SUM(PGP16+PGP21)/SUM(PGP13+PGP15+PGP20+PGP28+PGP31+PGP32)
PGP54	Ratio of fucosylated monosialylated structures with and without bisecting GlcNAc	PGP16/SUM(PGP13+PGP15)
PGP55	The incidence of bisecting GlcNAc in all fucosylated monosialylated structures	PGP16/SUM(PGP13+PGP15+PGP16)

PGP56	Ratio of fucosylated disialylated structures with and without bisecting GlcNAc	PGP21/SUM(PGP20+PGP24)
PGP57	The incidence of bisecting GlcNAc in all fucosylated disialylated structures	PGP21/SUM(PGP20+PGP21+PGP24)
PGP58	The percentage of FA2 in total neutral plasma glycans (GPn)	PGP1/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP59	The percentage of FA2B in total neutral plasma glycans (GPn)	PGP2/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP60	The percentage of A2BG1 in total neutral plasma glycans (GPn)	PGP3/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP61	The percentage of FA2G1 in total neutral plasma glycans (GPn)	PGP4/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP62	The percentage of FA2G1 in total neutral plasma glycans (GPn)	PGP5/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP63	The percentage of FA2BG1 in total neutral plasma glycans (GPn)	PGP6/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP64	The percentage of M6 in total neutral plasma glycans (GPn)	PGP7/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP65	The percentage of A2G2 in total neutral plasma glycans (GPn)	PGP8/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP66	The percentage of A2BG2 in total neutral plasma glycans (GPn)	PGP9/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP67	The percentage of FA2G2 in total neutral plasma glycans (GPn)	PGP10/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP68	The percentage of FA2BG2 in total neutral plasma glycans (GPn)	PGP11/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP69	The percentage of M9 in total neutral plasma glycans (GPn)	PGP18/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP70	The percentage of agalactosylated structures in total neutral plasma glycans	SUM(PGP1+PGP2+PGP7+1/2*PGP12+PGP18)/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP71	The percentage of monogalactosylated structures in total neutral plasma glycans	SUM(PGP3+PGP4+PGP5+PGP6)/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP72	The percentage of digalactosylated structures in total neutral plasma glycans	SUM(PGP8+PGP9+PGP10+PGP11)/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PG

		P7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP73	The percentage of all fucosylated structures (+/- bisecting GlcNAc) in total neutral plasma glycans	SUM(PGP1+PGP2+PGP4+PGP5+PGP6+PGP10+PGP11)/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP74	The percentage of fucosylation of monogalactosylated structures in total neutral plasma glycans	SUM(PGP4+PGP5+PGP6)/SUM(PGP3+PGP4+PGP5+PGP6) * 100
PGP75	The percentage of fucosylation of digalactosylated structures in total neutral plasma glycans	SUM(PGP10+PGP11)/SUM(PGP8+PGP9+PGP10+PGP11) * 100
PGP76	The percentage of fucosylated structures (without bisecting GlcNAc) in total neutral plasma glycans	SUM(PGP1+PGP4+PGP5+PGP10)/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP77	The percentage of fucosylation of agalactosylated structures (without bisecting GlcNAc) in total neutral plasma glycans	PGP1/SUM(PGP1+PGP2+PGP7+1/2*PGP12+PGP18) * 100
PGP78	The percentage of fucosylation of monogalactosylated structures (without bisecting GlcNAc) in total neutral plasma glycans	SUM(PGP4+PGP5)/SUM(PGP3+PGP4+PGP5+PGP6) * 100
PGP79	The percentage of fucosylation of digalactosylated structures (without bisecting GlcNAc) in total neutral plasma glycans	PGP10/SUM(PGP8+PGP9+PGP10+PGP11) * 100
PGP80	The percentage of fucosylated structures (with bisecting GlcNAc) in total neutral plasma glycans	SUM(PGP2+PGP6+PGP11)/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP81	The percentage of fucosylation of agalactosylated structures (with bisecting GlcNAc) in total neutral plasma glycans	PGP2/SUM(PGP1+PGP2+PGP7+1/2*PGP12+PGP18) * 100
PGP82	The percentage of fucosylation of monogalactosylated structures (with bisecting GlcNAc) in total neutral plasma glycans	PGP6/SUM(PGP3+PGP4+PGP5+PGP6) * 100
PGP83	The percentage of fucosylation of digalactosylated structures (with bisecting GlcNAc) in total neutral plasma glycans	PGP11/SUM(PGP8+PGP9+PGP10+PGP11) * 100
PGP84	Ratio of fucosylated structures with and without bisecting GlcNAc in total neutral plasma glycans	SUM(PGP2+PGP6+PGP11)/SUM(PGP1+PGP4+PGP5+PGP10)
PGP85	The incidence of bisecting GlcNAc in all fucosylated structures in total neutral plasma glycans	SUM(PGP2+PGP6+PGP11)/SUM(PGP1+PGP2+PGP4+PGP5+PGP6+PGP10+PGP11) * 100
PGP86	Ratio of fucosylated non-bisecting GlcNAc structures and all structures with	SUM(PGP1+PGP4+PGP5+PGP10)/SUM(PGP2+PGP3+PGP6+PGP9+PGP11)

	bisecting GlcNAc in total neutral plasma glycans	
PGP87	Ratio of afucosylated structures with bisecting GlcNAc and all fucosylated structures (+/- bisecting GlcNAc) in total neutral plasma glycans	SUM(PGP3+PGP9)/SUM(PGP1+PGP2+PGP4+PGP5+PGP6+PGP10+PGP11)
PGP88	Ratio of fucosylated digalactosylated structures with and without bisecting GlcNAc in total neutral plasma glycans	PGP11/PGP10
PGP89	The incidence of bisecting GlcNAc in all fucosylated digalactosylated structures in total neutral plasma glycans	PGP11/SUM(PGP10+PGP11) * 100
PGP90	Ratio of fucosylated digalactosylated non-bisecting GlcNAc structures and all digalactosylated structures with bisecting GlcNAc in total neutral plasma glycans	PGP10/SUM(PGP9+PGP11)
PGP91	Ratio of afucosylated digalactosylated structures with bisecting GlcNAc and all fucosylated digalactosylated structures (+/- bisecting GlcNAc) in total neutral plasma glycans	PGP9/SUM(PGP10+PGP11)
PGP92	The percentage of antennary fucosylated structures in total plasma glycans	PGP24+PGP30+PGP32+PGP36
PGP93	The percentage of core fucosylated structures in total plasma glycans	PGP1+PGP2+PGP4+PGP5+PGP6+PGP10+PGP11+PGP13+PGP15+PGP16+PGP20+PGP21+PGP28+PGP31+PGP32
PGP94	The percentage of neutral glycan structures in total plasma glycans	PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18
PGP95	the percentage of monosyalated structures in total plasma glycans	1/2*PGP12+PGP13+PGP14+PGP15+PGP16
PGP96	the percentage of bisyalated structures in total plasma glycans	PGP17+PGP19+PGP20+PGP21+PGP22+PGP23+PGP24
PGP97	the percentage of trisyalated structures in total plasma glycans	PGP25+PGP26+PGP27+PGP28+PGP29+PGP30+PGP31+PGP32+PGP33
PGP98	the percentage of tetrasyalated structures in total plasma glycans	PGP34+PGP35+PGP36
PGP99	The percentage of agalactosylated structures in total plasma glycans	PGP1+PGP2+PGP7+1/2*PGP12+PGP18
PGP100	The percentage of monogalactosylated structures in total plasma glycans	PGP3+PGP4+PGP5+PGP6+PGP13
PGP101	The percentage of digalactosylated structures in total plasma glycans	PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP14+PGP15+PGP16+PGP17+PGP19+PGP20+PGP21
PGP102	The percentage of trigalactosylated structures in total plasma glycans	PGP22+PGP23+PGP24+PGP25+PGP26+PGP27+PGP28+PGP29+PGP30+PGP31+PGP32
PGP103	The percentage of tetragalactosylated structures in total plasma glycans	PGP33+PGP34+PGP35+PGP36
PGP104	The percentage of biantennary structures in total plasma glycans	PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PG

		P13+PGP14+PGP15+PGP16+PGP17+PGP19+PGP20+PGP21
PGP105	The percentage of triantennary structures in total plasma glycans	PGP22+PGP23+PGP24+PGP25+PGP26+PGP27+PGP28+PGP29+PGP30+PGP31+PGP32
PGP106	The percentage of tetraantennary structures in total plasma glycans	PGP33+PGP34+PGP35+PGP36
PGP107	The percentage of high-mannose structures in total plasma glycans	PGP7+1/2*PGP12+PGP18
PGP108	The percentage of glycan structures with bisecting GlcNAc in total plasma glycans	PGP2+PGP3+PGP6+PGP9+PGP11+PGP16+PGP21
PGP109	Ratio of disialylated and trisialylated trigalactosylated structures	SUM(PGP22+PGP23+PGP24)/SUM(PGP25+PGP26+PGP27+PGP28+PGP29+PGP30+PGP31+PGP32)
PGP110	Ratio of trisialylated and tetrasialylated tetragalactosylated structures	PGP33/SUM(PGP34+PGP35+PGP36)
PGP111	The percentage of core-fucosylation of trigalactosylated structures	SUM(PGP28+PGP31+PGP32)/SUM(PGP22+PGP23+PGP24+PGP25+PGP26+PGP27+PGP28+PGP29+PGP30+PGP31+PGP32)*100
PGP112	The percentage of antennary-fucosylation of trigalactosylated structures	SUM(PGP24+PGP30+PGP32)/SUM(PGP22+PGP23+PGP24+PGP25+PGP26+PGP27+PGP28+PGP29+PGP30+PGP31+PGP32)*100
PGP113	The percentage of antennary-fucosylation of tetragalactosylated structures	PGP36/SUM(PGP33+PGP34+PGP35+PGP36)*100
PGP114	The percentage of M7 in total neutral plasma glycans (GPn)	1/2*PGP12/SUM(PGP1+PGP2+PGP3+PGP4+PGP5+PGP6+PGP7+PGP8+PGP9+PGP10+PGP11+1/2*PGP12+PGP18) * 100
PGP115	The percentage of syalated structures in total plasma glycans	1/2 * PGP12 + PGP13 + PGP14 + PGP15 + PGP16 + PGP17 + PGP19 + PGP20 + PGP21 + PGP22 + PGP23 + PGP24 + PGP25 + PGP26 + PGP27 + PGP28 + PGP29 + PGP30 + PGP31 + PGP32 + PGP33 + PGP34 + PGP35 + PGP36
PGP116	The percentage of galactosylated structures in total plasma glycans	PGP3 + PGP4 + PGP5 + PGP6 + PGP8 + PGP9 + PGP10 + PGP11 +1/2 * PGP12 + PGP13 + PGP14 + PGP15 + PGP16 + PGP17 + PGP19 + PGP20 + PGP21 + PGP22 + PGP23 + PGP24 + PGP25 + PGP26 + PGP27 + PGP28 + PGP29 + PGP30 + PGP31 + PGP32 + PGP33 + PGP34 + PGP35 + PGP36
PGP117	The percentage of galactosylated structures in total neutral plasma glycans	PGP3 + PGP4 + PGP5 + PGP6 + PGP8 + PGP9 + PGP10 + PGP11

Доп. табл. 4. Фактор инфляции тестовой статистики. Оценки получены для 113 признаков на основе результатов ПГИА, проведенного на материале выборки TwinsUK

Признак	Lambda GC	Признак	Lambda GC	Признак	Lambda GC
PGP1	1,0084	PGP39	0,9977	PGP77	1,0033
PGP2	1,0197	PGP40	1,0051	PGP78	1,0188
PGP3	0,9991	PGP41	1,0023	PGP79	1,0061
PGP4	1,0084	PGP42	1,0028	PGP80	1,0136
PGP5	1,0122	PGP43	1,0075	PGP81	1,0033
PGP6	1,0117	PGP44	1,0056	PGP82	1,0212
PGP7	1,0098	PGP45	1,0089	PGP83	1,0145
PGP8	0,9935	PGP46	1,0131	PGP84	1,0150
PGP9	1,0047	PGP47	1,0070	PGP85	1,0141
PGP10	1,0169	PGP48	1,0084	PGP86	1,0127
PGP11	1,0070	PGP49	1,0155	PGP87	1,0037
PGP12	0,9967	PGP50	1,0160	PGP88	1,0112
PGP13	1,0033	PGP51	1,0150	PGP89	1,0127
PGP14	1,0065	PGP52	1,0160	PGP90	1,0070
PGP15	1,0019	PGP53	1,0084	PGP91	0,9991
PGP16	1,0089	PGP54	1,0009	PGP92	1,0042
PGP17	1,0094	PGP55	1,0019	PGP93	1,0112
PGP18	1,0089	PGP56	1,0094	PGP94	1,0127
PGP19	1,0164	PGP57	1,0094	PGP95	1,0019
PGP20	1,0042	PGP58	1,0023	PGP96	1,0065
PGP21	0,9991	PGP59	1,0051	PGP97	0,9977
PGP22	0,9977	PGP60	1,0005	PGP98	1,0103
PGP23	1,0014	PGP61	1,0065	PGP99	1,0098
PGP24	1,0080	PGP62	1,0127	PGP100	1,0098
PGP25	1,0005	PGP63	1,0235	PGP101	1,0084
PGP26	1,0061	PGP64	0,9972	PGP102	1,0028
PGP27	0,9991	PGP65	1,0075	PGP103	1,0061
PGP28	1,0108	PGP66	0,9995	PGP104	1,0098
PGP29	1,0028	PGP67	1,0070	PGP105	1,0023
PGP30	1,0037	PGP68	0,9991	PGP106	1,0056
PGP31	1,0047	PGP69	1,0122	PGP107	1,0145
PGP32	1,0065	PGP70	1,0061	PGP108	1,0117
PGP33	1,0065	PGP71	1,0112	PGP109	1,0005
PGP34	1,0056	PGP72	1,0061	PGP110	1,0014
PGP35	1,0094	PGP73	1,0065	PGP111	1,0108
PGP36	1,0084	PGP74	1,0023	PGP112	1,0056
PGP37	1,0019	PGP75	1,0009	PGP113	1,0098
PGP38	1,0023	PGP76	1,0131		